Treating Dictionaries as a Linked-Data Corpus

Peter Bouda and Michael Cysouw

Abstract In this paper we describe a practical approach to the challenge of linguistic retrodigitization. We propose to distinguish strictly between a base digitization and separate interpretation of the sources. The base digitization only includes a literal electronic transcript of the source. All sources are thus simply treated as strings of characters, i.e. as unstructured corpora. The often complex structure as found in many dictionaries and grammars will subsequently (and possibly much later) be added as Linked Data in the form of standoff annotation. A further advantage of this approach is that the complete digitization and interpretation can be performed collaboratively without a complex organizational superstructure.

1 Introduction

A large amount of the knowledge about the world's languages is currently only available in traditionally printed form, as grammars, text collections or dictionaries. Although this body of knowledge is large, it is finite and manageable in size given current computational power and electronic storage. A proper retrodigitization of these resources would allow for many new approaches to the quantitative comparison of languages, be it for a better understanding of cross-linguistic variation in grammatical structure or for new and improved historical-comparative reconstructions.

Still, the number of pages to be digitized is large enough to pose serious challenges for the organizational infrastructure (we estimate the number of pages to be

Peter Bouda
Research Unit "Quantitative Language Comparison", Ludwig Maximilians University Munich, e-mail: pbouda@cidles.eu

Michael Cysouw
Research Unit "Quantitative Language Comparison", Ludwig Maximilians University Munich, e-mail: cysouw@lmu.de

digitized for the world's lesser-studied languages to be in the order of 10^6). Not only the size, but also the necessary precision of the digitization poses special desiderata. To allow for proper linguistic analysis, the precision of the digitization has to be highly accurate, because linguistic description is a strongly technical tradition in which each dot, dash, and tilde, and all italics, boldfaces and tab-marks have a specific and important meaning – and unfortunately a different meaning in each source.

This digitization will probably never be perfect the first time round. Also, the interpretation of all the special symbols used will be a task to be handled in many years to come, long after the basic digital encoding of the sources has been completed. The real challenge of linguistic retrodigitization is thus not the digital encoding as such, but the continuing update, refinement, and interpretation of the digital products.

In our view it is of central importance that everybody working with the digitized data should always able to trace back the information to the original source. Further, it should be possible to reconstruct every step in the digitization workflow to make it possible to find and correct errors. We propose a framework that allows scientists to work and enrich the digital data while maintaining this traceability. To accomplish this, we describe several technical and architectural solutions we devised in our project in which we are retrodigitizing dictionaries. What we create is a new type of linked-data corpus that is derived from legacy printed material and that generates new opportunities in research for a global scientific community, if done right.

2 Base Digitization and Annotation

The first step in the digitization process is the scanning and transcription of printed dictionaries. The end product of this transcription process is a basic text document with typographic and layout information. This transcript is transformed into an XML document which is the basis for the subsequent processing steps, which we describe in this paper. This raw digital version should minimally have basic formatting tags mimicking the printed original (i.e. italic, bold, etc), the original line breaks and indentation, and information about page and column numbers. From this information it is possible to approximately (though not necessarily perfectly) recreate the text as it looks in the original printed source.

Most importantly, this raw digital version does not include any interpretation about what the structure of the printed original is supposed to mean. For example, many dictionaries use italics to signify structure (e.g. parts of speech, or examples), but this structure will only be added later as Linked Data, so differing interpretations are possible. For such a interlinked structure to remain intact, the base version has to be as static and persistent as possible. So, we prefer a maximally simple base digitization as a start.

2.1 Basic Chunking

To ease annotation, the whole document is separated into reasonable and manageable chunks. Those chunks should be small enough to allow annotation through character counts. Although character counts could of course just as well work with complete books, for reasons of error correction and traceability we prefer chunks not larger than about a thousand characters (i.e. paragraph size). The chunks should also not be too small, so as to allow a human reader to quickly understand what she is looking at. Again, this is purely for reasons of manageability. In retrodigitization, we think that it is important not only to consider technical considerations, but also include arguments pertaining to social management and human interfaces. So we propose not to use word chunking, or even more complex linguistically-based chunking on the basic digitalization level. Further, also for reasons of traceability, those chunks should preferably be derived from the inherent structure of the sources. In our case of printed dictionaries we decided to use the entries of our dictionaries as basic tokens. For other sources, any available paragraph structure can be used to define chunks.

It is also necessary to remember the page number for each chunk, as we want to be able to approximately reconstruct the original printed pages. As the unique ID of each chunk, we use a human readable description which consists of two parts: the (start) page number and the relative position on the page. Table 1 shows one example entry and the information we store for it in the base digitization.

Table 1 A dictionary entry in the base digitization.

Field	Value
fullentry	afebeba (s1B) cuarto de arriba.
start page	22
start column	1
position on page	2

2.2 Adding Source Information as Annotations

We prefer a base digitization that does not include any internal structure except for the linear structure of the text, as this makes their handling much easier later on. So, all formatting information that is present in the original XML transcript is removed. This information is stored as Linked Data in the form of standoff annotations. The annotation refers to entries via its ID and by using character counts. For example, one annotation could have the information that the entry `/5/10/` (i.e. the one on page 5 at position 10) has italic characters from character position 9 to 14, another

annotation contains a newline at position 23. Table 2 shows one such annotation as an example.

Table 2 Data fields for annotation.

Field	Value
type	pagelayout
value	newline
start	23
end	23

Our annotation tools also alter the original transcript in one important way: we remove hyphens before line breaks and instead store an annotation called `hyphen` for the position where we removed it. Hyphenation is not considered to be content-related information, but only induced by the printed format of the original. We remove it from the base digitization because it is not necessary for later interpretation. For reasons of traceability, we keep the information, so we can reconstruct the original using this annotation.

2.3 Step-by-Step Enrichment

One of the advantages of this basic division into 'flat' base digitization and standoff annotations is the possibility to add information step-by-step by adding further annotations. For example, in our project the prime emphasis is on using the head words and translations of the dictionaries. We are able to find this information easily within the entries. Often, head words are printed in special format (bold or italic) and translations start and end after or before certain characters. We then save this induced information in the same way as the annotations mentioned in the previous paragraph, i.e. by adding exactly the same kind of standoff annotations.

Many of our dictionaries contain additional information about part-of-speech, some of them also have phonological and morphological descriptions of head words. There are often example sentences with translations, and all kind of further information. We do not parse all this information right now, as we do not need it for the current project. But other researchers can extract this information if wanted. In simple cases we already add additional annotations, but in other cases we leave this task to future research projects that might be interested in different information provided in the sources. The basic structure of our corpus allows us to focus on the things we need right now, but still open up the possibility of enriching the data in the future. No matter whether we do it ourselves or someone else adds interpretations.

3 (Re-)Publication and Collaboration

In addition to the internal work with the sources within our project, our goal is to publish the digitized dictionaries as a corpus. The data should be free to use by anyone (pending copyright issues). Again, there are two main principles our corpus structures needs to fulfill. On one hand, every researcher should be able to trace back everything we did with the data, up to the original entry, on the original page in the printed dictionary. Further, when a researcher thinks our annotations are not good enough, or wants to add information to our annotations, she should simply be able to do this, in an easy and independent way. Our framework proposes several means how to follow these principles, namely the usage of standoff annotations, an XML format and a certain URL structure to maintain traceability even all the way to using a printed URL.

3.1 Standoff Annotations

Storing information as standoff annotations has several advantages. First, users can just download the type of data they need. If someone needs plain dictionary entries, then she downloads the basic data file. If she needs additional information about line breaks and indentation, she downloads another file. This modularization makes data handling easier, even more so when more and more layers of annotations are added.

Second, the basic data has only the information one needs for automated linguistic analysis. It contains plain stings stripped of any structural information. Standards like those of the Text Encoding Initiative[1] (TEI) propose to store formatting, layout and structural information (especially for dictionaries) within the basic data. In our view, this leads to problems when later enriching the data with additional annotations. It is not very clear how tokenization and standoff annotation should work when tags are used extensively inside the basic data (cf. Cayless and Soroka, 2010; Bánski and Przepiórkowski, 2009).

Third, using a simple yet powerful and far-reaching standoff annotation from the beginning allows us to collaborate with other scientists, for example specialists for the language families we work with. Researchers can just add annotations without the need to remove or alter any of our annotations. It is also easily possible to integrate changes of the data with our annotations, if we want to permanently store them. But this integration is always optional: if we or someone else wants to publish personal interpretations of certain annotations she is free to do so. Here, the basic entry with a fixed ID serves as a reference point that connects all linked annotations.

[1] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html`

3.2 XML Format

XML data can sometimes be hard to handle, especially if you have large files and complex structure. There are cases where researchers prefer to have plain text files to process data (Schmidt, 2010). In our case, we prefer an XML structure that only has a minimum complexity, but still can represent every information we have. Given that our basic chunks are just strings, and annotations have only few data fields to store, the resulting XML is easily manageable. We are currently using a XML format that is derived from the proposals of the Corpus Encoding Standard[2] (CES), an early application of the TEI standard (see listing 1). We are aware that this standard is not actively developed anymore, and that the TEI is working on new standards that should also fulfill our needs in the end (Lee and Romary, 2010). But right now we see the CES-XML as the best way to store and exchange data like ours. CES is very easy to read, the specification is quite clear and focussed on the usability of data in different environments. Given the simple data structure of our work, any transformation into different XML structures should be trivial.

In general, though, the structure in which the data is stored is just a collection of linked-data entities, so everything is perfectly compatible with a more forward-looking RDF approach. For reasons of practical manageability we have decided not to use RDF as the underlying data model, but (for the moment) to rely on more traditional concepts like data tables with types and values. Conceptually, it is trivial to transform our data into an RDF representation, but the practical effort involved has kept us from providing such an access to our data right now. That will be done in the near future.

Listing 1 Example XML snippet for dictionary entry.

```
<div type="dictentry">
  <p id="22.2">
    afebeba (s1B) cuarto de arriba.
  </p>
</div>
<chunk from="22.2/0">
  <tok type="pagelayout" value="newline"
       from="22.2/23" to="22.2/23">
    <orth></orth>
  </tok>
</chunk>
```

3.3 URLs as Source Pointers

URLs are one of the most important means to publish and exchange scientific research nowadays, yet still most URLs give no hint on what kind of data is available behind them. In digital archives the URLs unfortunately often only con-

[2] http://www.cs.vassar.edu/CES/

tain numerical IDs. We want to use our URLs as references to a web page, but also to the original source. It should be usable for online needs, but also in printed publications. A reader who reads an article about our data should be able to take the dictionary from her shelf and look up the entry we discuss in the paper. As a result, our URLs contain a transparent string ID for the book, page numbers and the position on the page for entries and their annotations. A link to the entry page of a dictionary part of a book, for example, looks like this: `/source/thiesen1998/dictionary-25-339.html`. In this case, `thiesen1998` is our ID for Thiesen and Thiesen (1998), the dictionary part begins on page 25 and ends on page 339 of the book. The following two URLs reference smaller parts of it, for example all entries on page 25: `/source/thiesen1998/25/index.html`, or the 9th entry on page 25 and its annotations: `/source/thiesen1998/25/9/index.html`.

This structure will also be preserved when we transfer the data to any file-based archive. In a file-based archive, there will be a "main" folder for each dictionary (called `thiesen1998` in this case) and several sub-folders for pages and entries. This mapping of URLs to a files-and-folders structures and vice versa reduces the costs of data handling (as one can simply mirror our website to have a full archive structure) and allows easy traceability of every reference that will be published in the future, back to a web page or an archive folder, and even to the original dictionary.

3.4 Tags as Source Pointers

In addition to the URLs we propose another, more general possibility to refer to original sources and to cite in online publications. This technique is derived from tagging facilities in blogging and micro-blogging systems. The tags there are normally used to group the entries of one or more blogs under certain keywords, for example the tag `Linguistics` or hashtag `#linguistics` is used to group all blog articles, tweets, etc. that have linguistic content. We propose to adapt this procedure to allow scientist refer to the sources in an easy and intuitive way. The basic idea is to have a special tag (we propose `litref` to differentiate from existing tags) and add specifications about book, pages, URLS, etc. separated by slashes. If a scientist wants to refer to page 25 of the book (Thiesen and Thiesen, 1998) then a hashtag might look like this: `#litref/thiesen1998/page25`. To be more specific a ISBN or OCLC code may be used: `#litref/oclc/40505215`. The format of the tag should be as free as possible, as the most important thing is that scientists can cite in their electronic publications as easy as (or even easier than) in printed articles.

The next step is then to search and index those tags from all web pages, blogging and micro-blogging hosts. This task can be done by an adapted search robot that parses the structure of the tag and tries to find the source in a bibliographical database. Tags with ISBN, ISSN or some other codes are easy to parse, tags with references like `thiesen1998` might require some heuristics but should pose no bigger problem to state-of-the-art search robots. In the end the database consists of

the bibliographical entry plus all the web pages that refer to that entry and possibly additional information like page and line numbers. This results in a huge network of Linked Data that requires nothing more than users who agree on a certain tag and its template. Existing infrastructure like blogs and search robots can be used to create such a network.

Bibliographical entries are of course not restricted to printed publications. If researchers want to cite an electronic article (like a blog entry) they might just use the same tagging mechanism with a tag like `#litref/url/http://www...`; or even without the term `url` because it may be derived by the information given by the prefix `http://`. Another possible addition to the proposal is the introduction of a new (X)HTML meta-tag that contains all the tags for the sources that the current web page refers to. This is easy to do when a web page contains only one or two tags and makes it easier for search robot to harvest the tags and integrates this approach into the broader context of the semantic web. The whole infrastructure should still work without these meta-tags, it is an optional addition to the proposal.

A use case for such a framework in our project is a crowdsourcing approach for the corrections of dictionary entries and annotations. Possible co-workers may use a service like twitter to publish corrections to certain entries by using the proposed tags. A search robot than collects those tags and adds the content of the tweets to the entry in our database and on our website. We later manually apply the proposed changes to integrate the corrections into our database.

4 Summary

The combination of stable source URLs and the standoff annotation pointers provide a stable and easily manageable infrastructure for retrodigitization. As long as the source is kept simple and stable, multiple independent annotations can be added without the need for a central infrastructure, thus allowing collaborative annotation of the important and rich source of our linguistic heritage.

References

Bánski P, Przepiórkowski A (2009) Stand-off TEI annotation: The case of the National Corpus of Polish. In: Proceedings of the Third Linguistic Annotation Workshop (LAW III), pp 65–67

Cayless HA, Soroka A (2010) On implementing string-range() for TEI. In: Proceedings of Balisage: The Markup Conference 2010

Lee K, Romary L (2010) Towards interoperability of ISO standards for Language Resource Management. In: Proceedings of the 2nd International Conference on Global Interoperability for Language Resources

Schmidt D (2010) The inadequacy of embedded markup for cultural heritage texts. Literary and Linguistic Computing pp 337–356

Thiesen W, Thiesen E (1998) Diccionario Bora-Castellano Castellano-Bora. Instituto Lingüístico de Verano