Towards an ACL Anthology Corpus with Logical Document Structure

An Overview of the ACL 2012 Contributed Task

Ulrich Schäfer
DFKI Language Technology Lab
Campus D 3 1
D-66123 Saarbrücken, Germany
ulrich.schaefer@dfki.de

Jonathon Read, Stephan Oepen
Department of Informatics
Universitetet i Oslo
0316 Oslo, Norway
{jread|oe}@ifi.uio.no

Abstract

The ACL 2012 Contributed Task is a community effort aiming to provide the full ACL Anthology as a high-quality corpus with rich markup, following the TEI P5 guidelines—a new resource dubbed the ACL Anthology Corpus (AAC). The goal of the task is threefold: (a) to provide a shared resource for experimentation on scientific text; (b) to serve as a basis for advanced search over the ACL Anthology, based on textual content and citations; and, by combining the aforementioned goals, (c) to present a showcase of the benefits of natural language processing to a broader audience. The Contributed Task extends the current Anthology Reference Corpus (ARC) both in size, quality, and by aiming to provide tools that allow the corpus to be automatically extended with new content—be they scanned or born-digital.

1 Introduction—Motivation

The collection of the Association for Computational Linguistics (ACL) Anthology began in 2002, with 3,100 scanned and born-digital[1] PDF papers. Since then, the ACL Anthology has become *the* open access collection[2] of scientific papers in the area of Computational Linguistics and Language Technology. It contains conference and workshop proceedings and the journal *Computational Linguistics* (formerly the *American Journal of Computational Linguistics*). As of Spring 2012, the ACL Anthology comprises approximately 23,000 papers from 46 years.

Bird et al. (2008) started collecting not only the PDF documents, but also providing the textual content of the Anthology as a corpus, the *ACL Anthology Reference Corpus*[3] (ACL-ARC). This text version was generated fully automatically and in different formats (see Section 2.2 below), using off-the-shelf tools and yielding somewhat variable quality.

The main goal was to provide a *reference corpus* with fixed releases that researchers could use and refer to for comparison. In addition, the vision was formulated that manually corrected *ground-truth* subsets could be compiled. This is accomplished so far for citation links from paper to paper inside the Anthology for a controlled subset. The focus thus was laid on bibliographic and bibliometric research and resulted in the ACL Anthology Network (Radev et al., 2009) as a public, manually corrected citation database.

What is currently missing is an easy-to-process XML variant that contains high-quality running text and logical markup from the layout, such as section headings, captions, footnotes, italics etc. In principle this could be derived from LaTeX source files, but unfortunately, these are not available, and furthermore a considerable amount of papers have been typeset with various other word processing software.

Here is where the ACL 2012 Contributed Task starts: The idea is to combine OCR and PDFBox-like born-digital text extraction methods and reassign font and logical structure information as part of a rich XML format. The method would rely on OCR exclusively only in cases where no born-digital

[1] The term born-digital means natively digital, i.e. prepared electronically using typesetting systems like LaTeX, OpenOffice, and the like—as opposed to digitized (or scanned) documents.

[2] http://aclweb.org/anthology

[3] http://acl-arc.comp.nus.edu.sg

Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pages 88–97,
Jeju, Republic of Korea, 10 July 2012.

PDFs are available—in case of the ACL Anthology mostly papers published before the year 2000. Current results and status updates will always be accessible through the following address:

`http://www.delph-in.net/aac/`

We note that manually annotating the ACL Anthology is not viable. In a feasibility study we took a set of five eight-page papers. After extracting the text using PDFBox[4] we manually corrected the output and annotated it with basic document structure and cross-references; this took 16 person-hours, which would suggest a rough estimate of some 25 person-years to manually correct and annotate the current ACL Anthology. Furthermore, the ACL Anthology grows substantially every year, requiring a sustained effort.

2 State of Affairs to Date

In the following, we briefly review the current status of the ACL Anthology and some of its derivatives.

2.1 ACL Anthology

Papers in the current Anthology are in PDF format, either as scanned bitmaps or digitally typeset with LaTeX or word processing software. Older scanned papers were often created using type writers, and sometimes even contained hand-drawn graphics.

2.2 Anthology Reference Corpus (ACL-ARC)

In addition to the PDF documents, the ACL-ARC also contains (per page and per paper)

- bitmap files (in the PNG file format)
- plain text in 'normal' reading order
- formatted text (in two columns for most of the papers)
- XML raw layout format containing position information for each word, grouped in lines, with font information, but no running text variant.

The latter three have been generated using OCR software (OmniPage) operating on the bitmap files. However, OCR methods tend to introduce character and layout recognition errors, from both scanned and born-digital documents.

The born-digital subset of the ACL-ARC (mostly papers that appeared in 2000 or later) also contains PDFBox plain text output. However, this is not available for approximately 4% of the born-digital PDFs due to unusual font encodings. Note though, that extracting text from PDFs in normal reading order is not a trivial task (Berg et al., 2012), and many errors exist. Furthermore, the plain text is not dehyphenated, necessitating a language model or lexicon-based lookup for post-processing.

2.3 ACL Anthology Network

The ACL Anthology Network (Radev et al., 2009) is based on the ACL-ARC text outputs. It additionally contains manually-corrected citation graphs, author and affiliation data for most of the Anthology (papers until 2009).

2.4 Publications with the ACL Anthology as a Corpus

We did a little survey in the ACL Anthology of papers reporting on having used the ACL Anthology as corpus/dataset. The aim here is to get an overview and distribution of the different NLP research tasks that have been pursued using the ACL Anthology as dataset. There are probably other papers outside the Anthology itself, but these have not been looked at.

The pioneers working with the Anthology as corpus are Ritchie et al. (2006a, 2006b). They did work related to citations which also forms the largest topic cluster of papers applying or using Anthology data.

Later papers on citation analysis, summarization, classification, etc. are Qazvinian et al. (2010), Abu-Jbara & Radev (2011), Qazvinian & Radev (2010), Qazvinian & Radev (2008), Mohammad et al. (2009), Athar (2011), Schäfer & Kasterka (2010), and Dong & Schäfer (2011).

Text summarization research is performed in Qazvinian & Radev (2011) and Agarwal et al. (2011a, 2011b).

The HOO ("Help our own") text correction shared task (Dale & Kilgarriff, 2010; Zesch, 2011; Rozovskaya et al., 2011; Dahlmeier et al., 2011) aims at developing automated tools and techniques that

[4] `http://pdfbox.apache.org`

assist authors, e.g. non-native speakers of English, in writing (better) scientific publications.

Classification/Clustering related publications are Muthukrishnan et al. (2011) and Mao et al. (2010).

Keyword extraction and topic models based on Anthology data are addressed in Johri et al. (2011), Johri et al. (2010), Gupta & Manning (2011), Hall et al. (2008), Tu et al. (2010) and Daudaravičius (2012). Reiplinger et al. (2012) use the ACL Anthology to acquire and refine extraction patterns for the identification of glossary sentences.

In this workshop several authors have used the ACL Anthology to analyze the history of computational linguistics. Radev & Abu-Jbara (2012) examine research trends through the citing sentences in the ACL Anthology Network. Anderson et al. (2012) use the ACL Anthology to perform a people-centered analysis of the history of computational linguistics, tracking authors over topical subfields, identifying epochs and analyzing the evolution of subfields. Sim et al. (2012) use a citation analysis to identify the changing factions within the field. Vogel & Jurafsky (2012) use topic models to explore the research topics of men and women in the ACL Anthology Network. Gupta & Rosso (2012) look for evidence of text reuse in the ACL Anthology.

Most of these and related works would benefit from section (heading) information, and partly the approaches already used *ad hoc* solutions to gather this information from the existing plain text versions. Rich text markup (e.g. italics, tables) could also be used for linguistic, multilingual example extraction in the spirit of the ODIN project (Xia & Lewis, 2008; Xia et al., 2009).

3 Target Text Encoding

To select encoding elements we adopt the TEI P5 Guidelines (TEI Consortium, 2012). The TEI encoding scheme was developed with the intention of being applicable to all types of natural language, and facilitating the exchange of textual data among researchers across discipline. The guidelines are implemented in XML; we currently use inline markup, but stand-off annotations have also been applied (Bański & Przepiórkowski, 2009).

We use a subset of the TEI P5 Guidelines as not all elements were deemed necessary. This process was made easier through Roma[5], an online tool that assists in the development of TEI validators. We note that, while we initially use a simplified version, the schemas are readily extensible. For instance, Przepiórkowski (2009) demonstrates how constituent and dependency information can be encoded following the guidelines, in a manner which is similar to other prominent standards.

A TEI corpus is typically encoded as a single XML document, with several `text` elements, which in turn contain `front` (for abstracts), `body` and `back` elements (for acknowledgements and bibliographies). Then, sections are encoded using `div` elements (with `xml:ids`), which contain a heading (`head`) and are divided into paragraphs (`p`). We aim for accountability when translating between formats; for example, the `del` element records deletions (such as dehyphenation at line breaks).

An example of a TEI version of an ACL Anthology paper is depicted in Figure 1 on the next page.

4 An Overview of the Contributed Task

The goal of the ACL 2012 Contributed Task is to provide a high-quality version of the textual content of the ACL Anthology as a corpus. Its rich text XML markup will contain information on logical document structure such as section headings, footnotes, table and figure captions, bibliographic references, italics/emphasized text portions, non-latin scripts, etc.

The initial source are the PDF documents of the Anthology, processed with different text extraction methods and tools that output XML/HTML. The input to the task itself then consists of two XML formats:

- *PaperXML* from the ACL Anthology Searchbench[6] (Schäfer et al., 2011) provided by DFKI Saarbrücken, of all approximately 22,500 papers currently in the Anthology (except ROCLING which are mostly in Chinese). These were obtained by running a commercial OCR program and applying logical markup postprocessing and conversion to XML (Schäfer & Weitz, 2012).

[5] `http://www.tei-c.org/Roma/`

[6] `http://aclasb.dfki.de`

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.tei-c.org/ns/1.0 aclarc.tei.xsd" xml:lang="en">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Task-oriented Evaluation of Syntactic Parsers and Their Representations</title>
                <author>
                    Yusuke Miyao† Rune Sætre† Kenji Sagae† Takuya Matsuzaki† Jun'ichi Tsujii†‡*
                    †Department of Computer Science, University of Tokyo, Japan
                    ‡School of Computer Science, University of Manchester, UK
                    National Center for Text Mining, UK
                    {yusuke,rune.saetre,sagae,matuzaki,tsujii}@is.s.u-tokyo.ac.jp
                </author>
            </titleStmt>
            <publicationStmt>
                <publisher>Association for Computational Linguistics</publisher>
                <pubPlace> Columbus, Ohio, USA</pubPlace>
                <date>June 2008</date>
            </publicationStmt>
            <sourceDesc> [...] </sourceDesc>
        </fileDesc>
        <encodingDesc> [...] </encodingDesc>
    </teiHeader>
    <text>
        <front>
            <div type="abs">
                <head>Abstract</head>
                <p> [...] </p>
            </div>
        </front>
        <body>
            <div xml:id="SE1">
                <head>Introduction</head>
                <p>
                    Parsing technologies have improved considerably in
                    the past few years, and high-performance syntactic
                    parsers are no longer limited to PCFG-based frame<del type="lb">-</del>
                    works (<ref target="#BI6">Charniak, 2000</ref>;
                    [...]
                </p>
            </div>
        </body>
        <back>
            <div type="ack">
                <head>Acknowledgements</head>
                <p> [...] </p>
            </div>
            <div type="bib">
                <head>References</head>
                <listBibl>
                    <bibl xml:id="BI1">
                        D. M. Bikel. 2004. Intricacies of Collins' parsing model.
                        <hi rend="italic">Computational Linguistics</hi>, 30(4):479–511.
                    </bibl>
                    [...]
                </listBibl>
                <pb n="54"/>
            </div>
        </back>
    </text>
</TEI>
```

Figure 1: An example of a TEI-compliant version of an ACL Anthology document P08-1006. Some elements are truncated ([...]) for brevity.

- TEI P5 XML generated by PDFExtract. For papers from after 1999, an additional high-quality extraction step took place, applying state-of the art word boundary and layout recognition methods directly to the native, logical PDF structure (Berg et al., 2012). As no character recognition errors occur, this will form the master format for textual content if available.

Because both versions are not perfect, a large, initial part of the Contributed Task requires automatically adding missing or correcting markup, using information from OCR where necessary (e.g. for tables). Hence, for most papers from after 1999 (currently approx. 70% of the papers), the Contributed Task can make use of both representations simultaneously.

The role of *paperXML* in the Contributed Task is to serve as fall-back source (1) for older, scanned papers (mostly published before the year 2000), for which born-digital PDF sources are not available, or (2) for born-digital PDF papers on which the PDFExtract method failed, or (3) for document parts where PDFExtract does not output useful markup such as currently for tables, cf. Section 4.2 below.

A big advantage of PDFExtract is its ability to extract the full Unicode character range without character recognition errors, while the OCR-based extraction methods in our setup are basically limited to Latin1 characters to avoid higher recognition error rates.

We proposed the following eight areas as possible subtasks towards our goal.

4.1 Subtask 1: Footnotes

The first task addresses identification of footnotes, assigning footnote numbers and text, and generating markup for them in TEI P5 style. For example:

```
We first determine lexical heads of nonterminal
nodes by using Bikel's implementation of
Collins' head detection algorithm
<note place="foot" n="9">
 <hi rend="monospace">http://www.cis.upenn.edu/
 ~dbikel/software.html</hi>
</note>
(<ref target="#BI1">Bikel, 2004</ref>;
 <ref target="#BI11">Collins, 1997</ref>).
```

Footnotes are handled to some extent in PDFExtract and *paperXML*, but the results require refinement.

4.2 Subtask 2: Tables

Task 2 identifies figure/table references in running text and links them to their captions. The latter will also have to be distinguished from running text. Furthermore, tables will have to be identified and transformed into HTML style table markup. This is currently not generated by PDFExtract, but the OCR tool used for *paperXML* generation quite reliably recognizes tables and transforms tables into HTML. Thus, a preliminary solution would be to insert missing table content in PDFExtract output from the OCR results. In the long run, implementing table handling in PDFExtract would be desirable.

```
<ref target="#TA3">Table 3</ref> shows the
time for parsing the entire AImed corpus,...
<figure xml:id="TA3">
 <head>Table 3: Parsing time (sec.)</head>
 <!-- TEI table content markup here -->
</figure>
```

4.3 Subtask 3: Bibliographic Markup

The purpose of this task is to identify citations in text and link them to the bibliographic references listed at the end of each paper. In TEI markup, bibliographies are contained in `listBibl` elements. The contents of `listBibl` can range from formatted text to moderately-structured entries (`biblStruct`) and fully-structured entries (`biblFull`). For example:

```
We follow the PPI extraction method of
<ref target="#BI39">Sætre et al. (2007)</ref>,
which is based on SVMs ...
<div type="bib">
 <head>References</head>
 <listBibl>
  <bibl xml:id="BI39">
   R. Sætre, K. Sagae, and J. Tsujii. 2007.
   Syntactic features for protein-protein
   interaction extraction. In
   <hi rend="italic">LBM 2007 short papers</hi>.
  </bibl>
 </listBibl>
</div>
```

A citation extraction and linking tool that is known to deliver good results on ACL Anthology papers (and even comes with CRF models trained on this corpus) is ParsCit (Councill et al., 2008). In this volume, Nhat & Bysani (2012) provide an implementation for this task using ParsCit and discuss possible further improvements.

4.4 Subtask 4: De-hyphenation

Both *paperXML* and PDFExtract output contain soft hyphenation indicators at places where the original paper contained a line break with hyphenation. In *paperXML*, they are represented by the Unicode soft hyphen character (in contrast to normal dashes that also occur). PDFExtract marks hyphenation from the original text using a special element. However, both tools make errors: In some cases, the hyphens are in fact hard hyphens. The idea of this task is to combine both sources and possibly additional information, as in general the OCR program used for *paperXML* more aggressively proposes de-hyphenation than PDFExtract. Hyphenation in names often persists in *paperXML* and therefore remains a problem that will have to be addressed as well. For example:

```
In this paper, we present a comparative
eval<del type="lb">-</del>uation of syntactic
parsers and their output
represen<del type="lb">-</del>tations based on
different frameworks:
```

4.5 Subtask 5: Remove Garbage such as Leftovers from Figures

In both *paperXML* and PDFExtract output, text remains from figures, illustrations and diagrams. This occurs more frequently in *paperXML* than in PDFExtract output because text in bitmap figures undergoes OCR as well. The goal of this subtask is to recognize and remove such text.

Bitmaps in born-digital PDFs are embedded objects for PDFExtract and thus can be detected and encoded within TEI P5 markup and ignored in the text extraction process:

```
<figure xml:id="FI3">
 <graphic url="P08-1006/FI3.png" />
 <head>
  Figure 3: Predicate argument structure
 </head>
</figure>
```

4.6 Subtask 6: Generate TEI P5 Markup for Scanned Papers from *paperXML*

Due to the nature of the extraction process, PDFExtract output is not available for older, scanned papers. These are mostly papers from before 2000, but also e.g. EACL 2003 papers. On the other hand, *paperXML* versions exist for almost all papers of the ACL Anthology, generated from OCR output. They still need to be transformed to TEI P5, e.g. using XSLT. The *paperXML* format and transformation to TEI P5 is discussed in Schäfer & Weitz (2012) in this volume.

4.7 Subtask 7: Add Sentence Splitting Markup

Having a standard for sentence splitting with unique sentence IDs per paper to which everyone can refer to later could be important. The aim of this task is to add sentence segmentation to the target markup. It should be based on an open source tokenizer such as JTok, a customizable open source tool[7] that was also used for the ACL Anthology Searchbench semantic index pre-processing, or the Stanford Tokenizer[8].

```
<p><s>PPI extraction is an NLP task to identify
protein pairs that are mentioned as interacting
in biomedical papers.</s> <s>Because the number
of biomedical papers is growing rapidly, it is
impossible for biomedical researchers to read
all papers relevant to their research; thus,
there is an emerging need for reliable IE
technologies, such as PPI identification.
</s></p>
```

4.8 Subtask 8: Math Formulae

Many papers in the Computational Linguistics area, especially those dealing with statistical natural language processing, contain mathematical formulae. Neither *paperXML* nor PDFExtract currently provide a means to deal with these.

A math formula recognition is a complex task, inserting MathML[9] formula markup from an external tool (formula OCR, e.g. from InftyReader[10]) could be a viable solution.

For example, the following could become the target format of MathML embedded in TEI P5, for $\exists\delta > 0 \ni f(x) < 1$:

```
<mrow>
  <mo> there exists </mo>
  <mrow>
    <mrow>
      <mi> &#916; <!--GREEK SMALL DELTA--></mi>
      <mo> &gt; </mo>
      <mn> 0 </mn>
```

[7] http://heartofgold.opendfki.de/repos/trunk/jtok; LPGL license

[8] http://nlp.stanford.edu/software/tokenizer.shtml; GPL V2 license

[9] http://www.w3.org/TR/MathML/

[10] http://sciaccess.net/en/InftyReader/

```
    </mrow>
    <mo> such that </mo>
    <mrow>
      <mrow>
        <mi> f </mi>
        <mo> &#2061; <!--FUNCTION APPL.--></mo>
        <mrow>
          <mo> ( </mo>
          <mi> x </mi>
          <mo> ) </mo>
        </mrow>
      </mrow>
      <mo> &lt; </mo>
      <mn> 1 </mn>
    </mrow>
  </mrow>
</mrow>
```

An alternative way would be to implement math formula recognition directly in PDFExtract using methods known from math OCR, similar to the page layout recognition approach.

5 Discussion—Outlook

Through the ACL 2012 Contributed Task, we have taken a (small, some might say) step further towards the goal of a high-quality, rich-text version of the ACL Anthology as a corpus—making available both the original text and logical document structure.

Although many of the subtasks sketched above did not find volunteers in this round, the Contributed Task, in our view, is an on-going, long-term community endeavor. Results to date, if nothing else, confirm the general suitability of (a) using TEI P5 markup as a shared target representation and (b) exploiting the complementarity of OCR-based techniques (Schäfer & Weitz, 2012), on the one hand, and direct interpretation of born-digital PDF files (Berg et al., 2012), on the other hand. Combining these approaches has the potential to solve the venerable challenges that stem from inhomogeneous sources in the ACL Anthology—e.g. scanned, older papers and digital newer papers, generated from a broad variety of typesetting tools.

However, as of mid-2012 there still is no ready-to-use, high-quality corpus that could serve as a shared starting point for the range of Anthology-based NLP activities sketched in Section 1 above. In fact, we remain slightly ambivalent about our recommendations for utilizing the current state of affairs and expected next steps—as we would like to avoid much work getting underway with a version of the corpus that we know is unsatisfactory. Further, obviously, versioning and well-defined release cycles will be a prerequisite to making the corpus useful for comparable research, as discussed by Bird et al. (2008).

In a nutshell, we see two possible avenues forward. For the ACL 2012 Contributed Task, we collected various views on the corpus data (as well as some of the source code used in its production) in a unified SVN repository. Following the open-source, crowd-sourcing philosophy, one option would be to make this repository openly available to all interested parties for future development, possibly augmenting it with support infrastructure like, for example, a mailing list and shared wiki.

At the same time, our experience from the past months suggests that it is hard to reach sufficient momentum and critical mass to make substantial progress towards our long-term goals, while contributions are limited to loosely organized volunteer work. A possibility we believe might overcome these limitations would be an attempt at formalizing work in this spirit further, for example through a funded project (with endorsement and maybe financial support from organizations like the ACL, ICCL, AFNLP, ELRA, or LDC).

A potential, but not seriously contemplated 'business model' for the ACL Anthology Corpus could be that only groups providing also improved versions of the corpus would get access to it. This would contradict the community spirit and other demands, viz. that all code should be made publicly available (as open source) that is used to produce the rich-text XML for new papers added to the Anthology. To decide on the way forward, we will solicit comments and expressions of interest during ACL 2012, including of course from the R50 workshop audience and participants in the Contributed Task. Current results and status updates will always be accessible through the following address:

`http://www.delph-in.net/aac/`

The ACL publication process for conferences and workshops already today supports automated collection of metadata and uniform layout/branding. For future high-quality collections of papers in the area of Computational Linguistics, the ACL could think

about providing extended macro packages for conferences and journals that generate rich text and document structure preserving (TEI P5) XML versions as a side effect, in addition to PDF generation. Technically, it should be possible in both LaTeX and (for sure) in word processors such as OpenOffice or MS Word. It would help reducing errors induced by the tedious PDF-to-XML extraction this Contributed Task dealt with.

Finally, we do think that it will well be possible to apply the Contributed Task ideas and machinery to scientific publications in other areas, including the envisaged NLP research and existing NLP applications for search, terminology extraction, summarization, citation analysis, and more.

6 Acknowledgments

The authors would like to thank the ACL, the workshop organizer Rafael Banchs, the task contributors for their pioneering work, and the NUS group for their support. We are indebted to Rebecca Dridan for helpful feedback on this work.

The work of the first author has been funded by the German Federal Ministry of Education and Research, projects TAKE (FKZ 01IW08003) and Deependance (FKZ 01IW11003). The second and third authors are supported by the Norwegian Research Council through the VerdIKT programme.

References

Abu-Jbara, A., & Radev, D. (2011). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language techologies* (pp. 500–509). Portland, OR.

Agarwal, N., Reddy, R. S., Gvr, K., & Rosé, C. P. (2011a). Scisumm: A multi-document summarization system for scientific articles. In *Proceedings of the ACL-HLT 2011 system demonstrations* (pp. 115–120). Portland, OR.

Agarwal, N., Reddy, R. S., Gvr, K., & Rosé, C. P. (2011b). Towards multi-document summarization of scientific articles: Making interesting comparisons with SciSumm. In *Proceedings of the workshop on automatic summarization for different genres, media, and languages* (pp. 8–15). Portland, OR.

Anderson, A., McFarland, D., & Jurafsky, D. (2012). Towards a computational history of the ACL:1980–2008. In *Proceedings of the ACL-2012 main conference workshop: Rediscovering 50 years of discoveries.* Jeju, Republic of Korea.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 student session* (pp. 81–87). Portland, OR.

Bański, P., & Przepiórkowski, A. (2009). Stand-off TEI annotation: the case of the National Corpus of Polish. In *Proceedings of the third linguistic annotation workshop* (pp. 64–67). Suntec, Singapore.

Berg, Ø. R., Oepen, S., & Read, J. (2012). Towards high-quality text stream extraction from PDF. Technical background to the ACL 2012 Contributed Task. In *Proceedings of the ACL-2012 main conference workshop on Rediscovering 50 Years of Discoveries.* Jeju, Republic of Korea.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., & Tan, Y. F. (2008). The ACL Anthology Reference Corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the sixth international conference on language resources and evaluation (LREC-08).* Marrakech, Morocco.

Councill, I. G., Giles, C. L., & Kan, M.-Y. (2008). ParsCit: An open-source CRF reference string parsing package. In *Proceedings of LREC-2008* (pp. 661–667). Marrakesh, Morocco.

Dahlmeier, D., Ng, H. T., & Tran, T. P. (2011). NUS at the HOO 2011 pilot shared task. In *Proceedings of the generation challenges session at the 13th european workshop on natural language generation* (pp. 257–259). Nancy, France.

Dale, R., & Kilgarriff, A. (2010). Helping Our Own: Text massaging for computational linguistics as a new shared task. In *Proceedings of the 6th international natural language generation conference.* Trim, Co. Meath, Ireland.

Daudaravičius, V. (2012). Applying collocation segmentation to the ACL Anthology Reference Corpus. In *Proceedings of the ACL-2012 main conference workshop: Rediscovering 50 years of discoveries.* Jeju, Republic of Korea.

Dong, C., & Schäfer, U. (2011). Ensemble-style self-training on citation classification. In *Proceedings of 5th international joint conference on natural language processing* (pp. 623–631). Chiang Mai, Thailand.

Gupta, P., & Rosso, P. (2012). Text reuse with ACL: (upward) trends. In *Proceedings of the ACL-2012 main conference workshop: Rediscovering 50 years of discoveries.* Jeju, Republic of Korea.

Gupta, S., & Manning, C. (2011). Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th international joint conference on natural language processing* (pp. 1–9). Chiang Mai, Thailand.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 363–371). Honolulu, Hawaii.

Johri, N., Ramage, D., McFarland, D., & Jurafsky, D. (2011). A study of academic collaborations in computational linguistics using a latent mixture of authors model. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 124–132). Portland, OR.

Johri, N., Roth, D., & Tu, Y. (2010). Experts' retrieval with multiword-enhanced author topic model. In *Proceedings of the NAACL HLT 2010 workshop on semantic search* (pp. 10–18). Los Angeles, California.

Mao, Y., Balasubramanian, K., & Lebanon, G. (2010). Dimensionality reduction for text using domain knowledge. In *COLING 2010: Posters* (pp. 801–809). Beijing, China.

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., & Zajic, D. (2009). Using citations to generate surveys of scientific paradigms. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics* (pp. 584–592). Boulder, Colorado.

Muthukrishnan, P., Radev, D., & Mei, Q. (2011). Simultaneous similarity learning and feature-weight learning for document clustering. In *Proceedings of textgraphs-6: Graph-based methods for natural language processing* (pp. 42–50). Portland, OR.

Nhat, H. D. H., & Bysani, P. (2012). Linking citations to their bibliographic references. In *Proceedings of the ACL-2012 main conference workshop: Rediscovering 50 years of discoveries.* Jeju, Republic of Korea.

Przepiórkowski, A. (2009). TEI P5 as an XML standard for treebank encoding. In *Proceedings of the eighth international workshop on treebanks and linguistic theories* (pp. 149–160). Milano, Italy.

Qazvinian, V., & Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd international conference on computational linguistics (COLING 2008)* (pp. 689–696). Manchester, UK.

Qazvinian, V., & Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 555–564). Uppsala, Sweden.

Qazvinian, V., & Radev, D. R. (2011). Learning from collective human behavior to introduce diversity in lexical choice. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language techologies* (pp. 1098–1108). Portland, OR.

Qazvinian, V., Radev, D. R., & Ozgur, A. (2010). Citation summarization through keyphrase extraction. In *Proceedings of the 23rd international conference on computational linguistics (COLING 2010)* (pp. 895–903). Beijing, China.

Radev, D., & Abu-Jbara, A. (2012). Rediscovering ACL discoveries through the lens of ACL Anthology Network citing sentences. In *Proceedings of*

the ACL-2012 main conference workshop: Rediscovering 50 years of discoveries. Jeju, Republic of Korea.

Radev, D., Muthukrishnan, P., & Qazvinian, V. (2009). The ACL Anthology Network corpus. In *Proceedings of the 2009 workshop on text and citation analysis for scholarly digital libraries.* Morristown, NJ, USA.

Radev, D. R., Muthukrishnan, P., & Qazvinian, V. (2009). The ACL Anthology Network. In *Proceedings of the 2009 workshop on text and citation analysis for scholarly digital libraries* (pp. 54–61). Suntec City, Singapore.

Reiplinger, M., Schäfer, U., & Wolska, M. (2012). Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis. In *Proceedings of the ACL-2012 main conference workshop: Rediscovering 50 years of discoveries.* Jeju, Republic of Korea.

Ritchie, A., Teufel, S., & Robertson, S. (2006a). Creating a test collection for citation-based IR experiments. In *Proceedings of the human language technology conference of the NAACL, main conference* (pp. 391–398). New York City.

Ritchie, A., Teufel, S., & Robertson, S. (2006b). How to find better index terms through citations. In *Proceedings of the workshop on how can computational linguistics improve information retrieval?* (pp. 25–32). Sydney, Australia.

Rozovskaya, A., Sammons, M., Gioja, J., & Roth, D. (2011). University of illinois system in HOO text correction shared task. In *Proceedings of the generation challenges session at the 13th european workshop on natural language generation* (pp. 263–266). Nancy, France.

Schäfer, U., & Kasterka, U. (2010). Scientific authoring support: A tool to navigate in typed citation graphs. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics and writing: Writing processes and authoring aids* (pp. 7–14). Los Angeles, CA.

Schäfer, U., Kiefer, B., Spurk, C., Steffen, J., & Wang, R. (2011). The ACL Anthology Searchbench. In *Proceedings of the ACL-HLT 2011 system demonstrations* (pp. 7–13). Portland, OR.

Schäfer, U., & Weitz, B. (2012). Combining OCR outputs for logical document structure markup. Technical background to the ACL 2012 Contributed Task. In *Proceedings of the ACL-2012 main conference workshop on Rediscovering 50 Years of Discoveries.* Jeju, Republic of Korea.

Sim, Y., Smith, N. A., & Smith, D. A. (2012). Discovering factions in the computational linguistics community. In *Proceedings of the ACL-2012 main conference workshop: Rediscovering 50 years of discoveries.* Jeju, Republic of Korea.

TEI Consortium. (2012, February). *TEI P5: Guidelines for electronic text encoding and interchange.* (http://www.tei-c.org/Guidelines/P5)

Tu, Y., Johri, N., Roth, D., & Hockenmaier, J. (2010). Citation author topic model in expert search. In *COLING 2010: Posters* (pp. 1265–1273). Beijing, China.

Vogel, A., & Jurafsky, D. (2012). He said, she said: Gender in the ACL anthology. In *Proceedings of the ACL-2012 main conference workshop: Rediscovering 50 years of discoveries.* Jeju, Republic of Korea.

Xia, F., Lewis, W., & Poon, H. (2009). Language ID in the context of harvesting language data off the web. In *Proceedings of the 12th conference of the european chapter of the ACL (EACL 2009)* (pp. 870–878). Athens, Greece.

Xia, F., & Lewis, W. D. (2008). Repurposing theoretical linguistic data for tool development and search. In *Proceedings of the third international joint conference on natural language processing: Volume-i* (pp. 529–536). Hyderabad, India.

Zesch, T. (2011). Helping Our Own 2011: UKP lab system description. In *Proceedings of the generation challenges session at the 13th european workshop on natural language generation* (pp. 260–262). Nancy, France.