LexBib: A Corpus and Bibliography of Metalexicographical Publications

David Lindemann, Fritz Kliche, Ulrich Heid
Universität Hildesheim
E-mail: david.lindemann@uni-hildesheim.de, fritz.kliche@uni-hildesheim.de, heid@uni-hildesheim.de

Abstract

This paper presents preliminary considerations regarding objectives and workflow of LexBib, a project which is currently being developed at the University of Hildesheim. We briefly describe the state of the art in electronic bibliographies in general, and bibliographies of lexicography and dictionary research in particular. The LexBib project is intended to provide a collection of full texts and metadata of publications on metalexicography, as an online resource and research infrastructure; at the same time, LexBib has a strong experimental component: computational linguistic methods for automated keyword indexing, topic clustering and citation extraction will be tested and evaluated. The goal is to enrich the bibliography with the results of the text analytics in the form of additional metadata.

Keywords: bibliography, metalexicography, full text collection, e-science corpus, text analytics

1 Introduction

Domain-specific bibliographies are important tools for scientific research. We believe that much of their usefulness depends on the metadata they provide for (collections of) publications, and on advanced search functionalities. What is more, bibliographies for a limited domain may offer hand-validated publication metadata. As for lexicography and dictionary research, several bibliographies with different scopes and formats exist independently from each other; none of them covers the field completely, and most of them do not support advanced search functionalities, so that usability is dramatically reduced. Searches for bibliographical data and for the corresponding full texts are therefore most often performed using general search engines and domain-independent bibliography portals. However, big domain-independent repositories have two major shortcomings: They often contain noisy or incomplete publication metadata which have to be hand-validated by the users when copying them into their personal bibliographies, e. g. for citations. Closely related to that, the search functions of leading bibliography portals still focus on query-based information retrieval, since a combination of cascaded filter options using keywords and metadata such as persons, places, events, and relations to other items, only yields good results if the metadata meet certain requirements on precision and completeness.

Our goal is a domain-specific online bibliography of lexicography and dictionary research (i.e. metalexicography) which offers hand-validated publication metadata as they are needed for citations, and which in addition is complemented with the output of an NLP toolchain.

Several methods from computational linguistics produce useful results for seeking and retrieving scientific publications. For example, topic clustering has become very popular in the Digital Humanities. We suggest that assigning topics to publications provides valuable metadata for finding related work. Methods for term extraction have a similar objective. They detect text patterns (thus: terms) that are more significant in a (more specific) domain corpus than in a (more general) reference corpus.

Scientific publications usually contain a reference section. The analysis of citations is useful for the retrieval process in different dimensions. The number of citations a paper receives is an indicator of its scientific impact. Next, a citation network discloses clusters of collaborating researchers and of related work. Third, metadata on citations can be combined with other metadata in different ways; this is useful, for instance, when citation clusters are not strongly interconnected, but the corresponding authors still work on similar topics. Tools for parsing the reference sections of scientific publications (e. g. GROBID, Romary & Lopez 2015) use NLP methods because the high number of different citation styles makes the use of machine learning on text data desirable.

Section 2 discusses existing resources for lexicography and metalexicography. Section 3 details the goals of the LexBib project. Section 4 describes the NLP methods we use for providing the bibliographical items with additional metadata. In Section 5, we present some results of a study on overlaps between Lexicography and Digital Humanities, for which we have compiled the actual LexBib publication metadata and full text corpus, together with a similar collection of Digital Humanities publications.

2 The State of the Art

In the following subsections, we describe existing collections of full texts and/or metadata from the fields of lexicography and metalexicography in terms of scope and qualitative features, as well as the state of the art regarding online presentations of bibliographical databases.

2.1 Full Text and Metadata Collections of Metalexicographical Publications

For lexicography and research on dictionaries, some collections exist as printed publications or are accessible online. Table 1 contains a selective list of recently published bibliographies of (meta-) lexicography, and the list of publication metadata for the LexBib test set (see Table 3 in Section 5); for collecting publication metadata for LexBib, we focus on these resources in the first place, and also collect the corresponding full texts. Later we shall include the contents of further bibliographical data collections we might have access to, and search by ourselves for full texts and publication metadata. For the retrieval of relevant publications that have not been included in any of the existing bibliographies, we might use keywords and citation metadata extracted from our lexicography e-science corpus (*cf.* workflow description in Section 4.)

In addition to the metadata listed in Table 1, the resources differ in terms of the item types of the publications they contain. Only three resources are dedicated to dictionaries (domain: “Lex”). Regarding metalexicography (domain: “Metalex”), all resources cover scientific publications of any type (monographs, journal articles, articles in conference proceedings, book chapters, dissertations) as well as references to their containers (collective volumes such as handbooks, conference proceedings, etc.); some resources also contain references to other bibliographies. Córdoba Rodríguez’ collection is the only resource which includes relevant newspaper articles. While our LexBib test set only contains contributions in English, all other resources list articles in multiple languages; Ahumada focuses on Hispanic metalexicography which is represented in publications written mainly in Spanish.

Another feature to look at is whether the bibliographies present their contents as alphabetically ordered list or, additionally, in a thematic order. Córdoba Rodríguez groups the bibliographical data in hierarchically organized thematic blocks; EURALEX, in turn, presents its references according to approximately 125 different keywords. The items of Obelex Meta are manually keyword-indexed; these approximately 70 keywords function as filter option in the extended search interface.

Table 1: Some existing bibliographies of metalexicography.

Title	*Scope (years)*	*Scope (domains)*	*Scope (languages)*	*# Items*	*Format*
LexBib Testset	2000-2017	Metalex	English	2,056	Structured database
EURALEX Bibliography[1]	1600-2010	Lex/Metalex	Multiple	1,325	Unstructured list (pub. as Wiki)
Obelex Meta[2]	1982-2017	Metalex	Multiple	ca. 2,000	Structured database
WLWF[3]	1420-2016	Lex/Metalex	Multiple	2,370	Unstructured list (pub. as PDF)
Wiegand[4]	1850-2014	Lex/Metalex	Multiple	33,339	Unstructured list (pub. as PDF)
Hartmann[5]	1930-2007	Metalex	Multiple	ca. 570	Unstructured list (pub. as PDF)
Córdoba Rodríguez[6]	1940-2003	Metalex	Multiple	10,192	Structured database
Ahumada[7]	1535-2010	Metalex	Mainly Spanish	6,560	Structured database (in progress)

Obelex Meta and the LexBib test set are available to us as structured data collections stored in relational databases. All metadata are stored as attribute-value pairs, which is a necessary condition for their processing, e.g. by algorithms for duplicate merging, or for its representation in machine-readable formats such as BibTeX or TEI-XML.

Concerning the application of computational text analysis to metalexicographical full text collections, the lexicography community can already count on several studies that show the usefulness of this kind of methodology, including term extraction and bibliometrics, for depicting trends in our discipline (De Schryver 2009, 2012; Lew & De Schryver 2014).

2.2 Features of Bibliographical Databases as Online Resources

As an example of a state-of-the-art bibliographical database we may cite DBLP, an online bibliography of computer science[8] maintained at Trier University (Ley 2002; Weber et al. 2006). Features of DBLP relevant as a guiding reference for a resource like LexBib are its data model which includes indices for journals and conferences, TOC (table of contents) pages for single volumes, the disambiguated person index and individual author pages, and the data presentation, that, in addition to query-based access, allows multi-layered browsing and faceted search. Third, all DBLP bibliographic records are accessible in multiple formats via an API, so that personal reference managers (e. g. *Zotero*) can take advantage of downloading metadata sets individually as well as in bulk. As a fourth point we may add that advanced search and visualization tools exist that use data retrieved from the DBLP API (Burch et al. 2015), and that could be fed with bibliographical data compliant to that format. A fifth guiding feature of DBLP that also matches to LexBib is its limited and well-defined scope as a specialized bibliography for one discipline. This is a condition for a resource that should stay small enough to be maintained noise-free by manual validation, and it limits the problem of irrelevant results in information retrieval, two problems that doubtlessly reduce the usability of global academic search engines.

[1] The EURALEX bibliography is accessible at http://euralex.pbworks.com.

[2] See Möhrs (2016). Accessible at http://www.owid.de/obelex/meta.

[3] Bibliography accessible to the editors of WLWF (Wiegand et al. 2010; 2017).

[4] Wiegand (2006–2015): 'Internationale Bibliographie'.

[5] Accessible at http://euralex.pbworks.com/f/Hartmann+Bibliography+of+Lexicography.pdf.

[6] Accessible at http://www.udc.es/grupos/lexicografia/bibliografia/index.html.

[7] I. Ahumada (ed.) (2006–2014): *Diccionario bibliográfico de la metalexicografía del español*. Starting with Volume III (2006–2010) and backwards, this work is being transformed into a structured database (*cf*. Porta Zamorano 2016).

[8] DBLP is accessible at http://dblp.dagstuhl.de.

In addition to unique identifiers and to standard publication metadata, i.e. the bibliographic data necessary for citing or referencing, some online repositories have started to perform citation extraction and semantic annotation of items using computational text analysis, and to provide the results as metadata for display and advanced search options (*cf.* e.g. the discussion in Zeni et al. 2007). In a future stage of the LexBib project, we aim at generating metadata of this kind, using an e-science corpus consisting of abstracts and full texts of publications in the domain of metalexicography as showcase.

3 Goals

The goals of the LexBib project can thus be described in an infrastructural and in a research dimension. On the one hand, it is our aim to provide an online bibliography of (meta-)lexicography that meets with the state of the art as described in Section 2.2. On the other hand, we will set up, test and evaluate a pipeline of NLP tools for citation extraction, and automatic keyword indexing, and it is our intention to include the results in the published version of the LexBib collection, marked as automatically generated publication metadata.

In general terms, the tasks which a user of an online bibliography might want to perform, and that the metadata-based searches we want to offer in LexBib shall consider, may be the following, among others (list adapted from Buch et al. 2015: 163):

- Papers with certain words or substrings in their titles, abstracts, and/or text bodies;
- Papers published in certain time frames, by certain persons, published in certain journals or presented at certain events;
- Keywords relevant for a specific time interval, list of authors, and subset of the bibliography (e. g. a conference series or an event);
- Keywords co-occurring with other words (multiword term candidates);
- Frequency distributions of correlated keywords presented in their distribution over time;
- Keyword correlations and citation relations between several authors;
- Author correlations and their change over time.

To this end, interactive (browsable) visualizations of keyword and author relations shall be created and made accessible as part of the LexBib online resource.

As for publication metadata to be included in LexBib, the intended minimal coverage for each item includes all metadata necessary for citing (*cf.* Section 4.1), as well as unique identifiers of publications (ISBN, DOI) and persons (ORCID), and item relations such as “is review of” and “is reviewed in” for reviews, “is part of” and “contains” for volumes, and “citing” and “is cited by”, regarding citations.

We are aware that the intended manual validation of publication metadata is a labor-intensive task, and we foresee a considerable amount of manual editing work, which we will track in detail in order to draw conclusions on how much manual work is necessary for a noise-free collection of bibliographical data. This kind of process metadata evaluation on the relatively limited LexBib e-science corpus may yield valuable hints for possible applications of the proposed workflow to larger e-science corpora. In a first phase, we propose to consider only items in English published between 2000 and 2017, and to move on towards other languages after an intermediate evaluation of the workflow, and later back to the past.

4 Methods

LexBib documents are provided with additional metadata on two dimensions: Publication metadata and metadata on the contents. Publication metadata are collected together with the full texts by semi-automatic means using the Zotero tool,[9] and they are manually validated (see Section 4.1). For retrieving contents metadata, PDF or HTML full texts are processed with the NLP pipeline described in Section 4.2.

4.1 Publication Metadata

As publication metadata, a predefined minimal metadata set is collected and hand-validated for every publication, including author, title, publishing year, name of the publication (e.g. the journal), place, DOI/ISBN, etc. Publication metadata include authors, their affiliations, the title of the publication, as well as document metadata like the source or the publishing year, which can be easily retrieved. The metadata of the LexBib test set collection that exists since 2018 will be merged with data from the resources listed in Table 1, as soon as they are available in or have been converted into a structured format (e.g. TEI-XML or BibTeX), in order (1) to obtain the intersecting set, i.e. duplicate items, for semi-automated metadata validation, (2) to enrich LexBib, and (3) to allow cross-resource referencing, i.e. to be able to point exactly to where an item appears in another bibliography.[10]

Regarding the use of some of the resources to be merged, practical and licensing issues will have to be addressed. In case an enrichment of LexBib will not be possible because of licensing issues, only cross-resource referencing is planned; nevertheless, also for that purpose, the publication metadata items to be referenced have to be accessible in a structured format.

4.2 Full Text Processing and Content Metadata

Full texts are cleaned and processed in the following way (see pipeline schema in Figure 1): Both PDF and HTML files are converted into plain text. The full text bodies are isolated, processed with the TreeTagger (Schmid, 1994) for part-of-speech tagging and lemmatization, which makes them accessible to topic clustering on lemmas and to term extraction. For citation extraction, the list of bibliographic references at the end of each full text is isolated and parsed (see Section 4.22).

4.2.1 Term Extraction and Topic Clustering

The full text content is converted into a lemmatized variant and processed with Mallet (McCallum 2002) for topic clustering. For each publication, Mallet provides a measure of how it is related to the different topics. Our goal is to use these assignments as LexBib metadata. The idea is to provide an access structure towards the items in the bibliography by browsing topics.

For term extraction, we use a tool suite developed at IMS (University of Stuttgart, cf. Rösiger et al. 2015; 2016). It extracts the instances of part-of-speech patterns, e. g. (1) NN (single common nouns), (2) NN-NN (two common nouns), or (3) NN-NN-NN (three adjacent common nouns). Then, it ranks the extracted instances according to their *termhood* or *keyness* which is measured by dividing the relative frequency of the instance in a text by the relative frequency of the instance in a reference corpus (*weirdness ratio*, cf. Ahmad et al. 1992). We run this method twice for each document; once with the British National Corpus (BNC) as a reference corpus in general language in order to retrieve

[9] See http://zotero.org.

[10] This can be useful, for example, if an item carries further information in the other resource, e.g. a short review, as in Wiegand (2006-2015).

domain specific terms; and once with the whole LexBib corpus as a reference corpus in order to identify text specific keywords. As an example of this procedure, the terms extracted from a Euralex 2014 keynote speech (Heid 2014) are given in Table 2. In the left column, they are ordered according to their keyness relative to the BNC, in the right column according to their keyness in comparison to the LexBib full text test set. It can be observed that terms relevant to Lexicography in general are ranked lower in the right column.

Table 2: Term candidates extracted from an example publication.

Top 20 Terms (Ref. BNC)	*Top 20 Terms (Ref. LexBib)*
text reception	information-on-demand
text production	on-demand
dictionary function	data repository
multiword	user friendliness
word formation	user orientation
production dictionary	text reception
user orientation	production dictionary
information-on-demand	text production
data repository	dictionary function
dictionary entry	repository
user friendliness	valency
internet	guidance
markup	orientation
on-demand	concord
valency	word formation
concord	scenario
dictionary	markup
corpus	production
collocation	classification
language processing	advance

In addition to manual revision and assessment of term candidates, we plan the evaluation of the term extraction results using the manually defined keywords given in Obelex Meta (Möhrs 2016) as a silver standard, in order to obtain data for adapting the performance of the automatic keyword indexing methods. A further possible application of the latter is a revision of the set of keywords used for indexing Obelex Meta items, and a grounding of term variants, i.e. an association of different variants of a term to a single keyword regarded as the canonical form or base variant (see discussion and methodology in Theofilidis 2018).

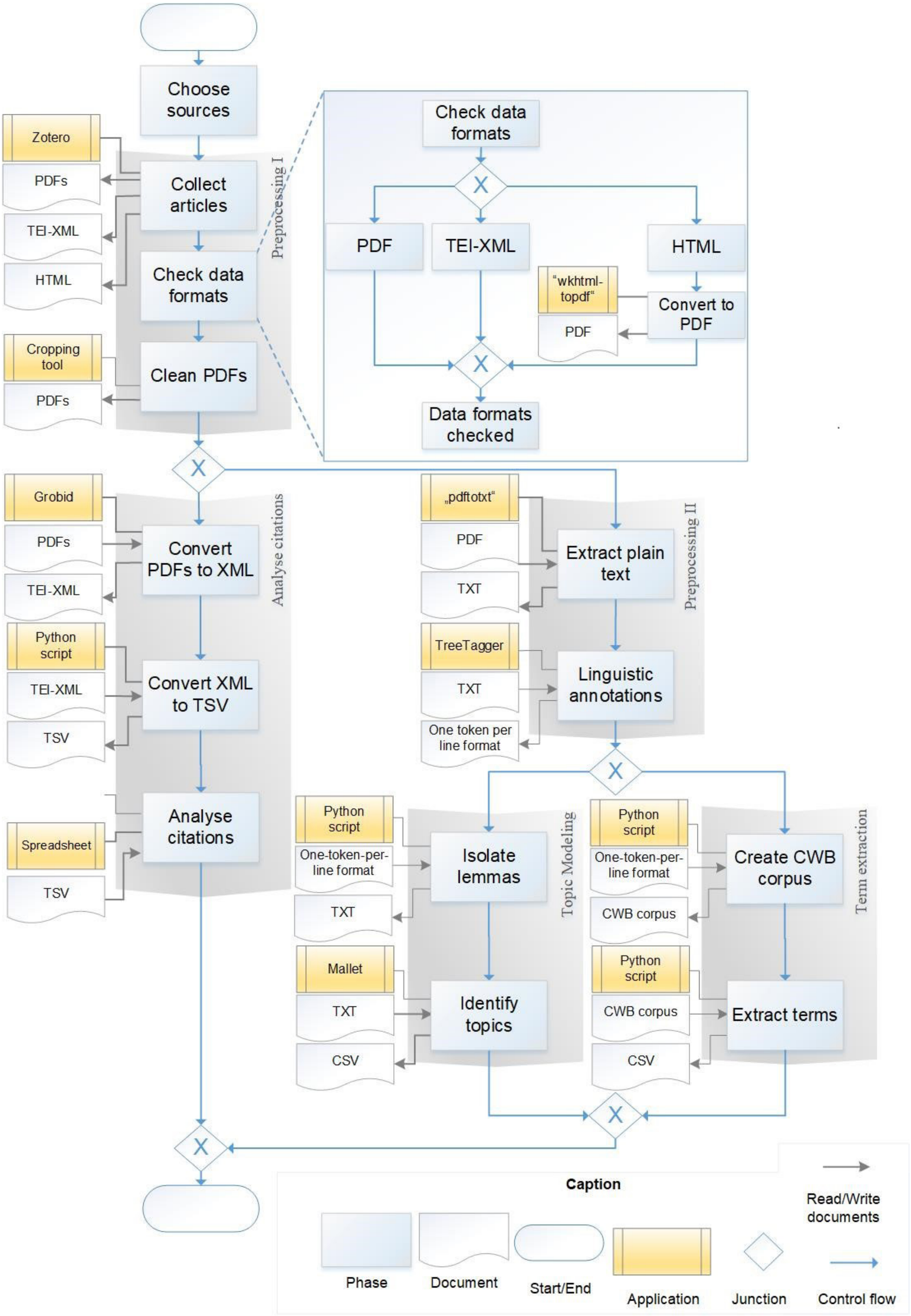

Figure 1: Workflow for LexBib corpus building and nlp processing.

4.2.2 Citation Network

The item relations obtained from the analysis of the reference sections in the full texts include (1) the publications cited in a publication, (2) the publications citing a publication, and (3) the membership of a publication in a cluster of a citation network. The GROBID tool (Lopez 2009) extracts a plain text version of the full text content and isolates the block of bibliographic references, the entries of which are then parsed and converted into a structured format compliant to the TEI guidelines (element <listBibl>). GROBID uses conditional random fields, a supervised machine learning method which learns a model based on annotated training data. Problematic citation styles, i.e. formats that are not properly parsed by the tool, will require further annotated training data. As a by-product, GROBID's recognition performance will be enhanced.

Based on the extracted references and the publication metadata sets, a citation network is modeled and publication clusters are identified. The analysis requires a mapping from a citation given in a publication to the metadata of the cited publication. We are aware that the metadata given in citations differ significantly. Letters with diacritics may be replaced with those without diacritics. The titles can also differ, e.g. subtitles can be left out. In our preliminary study (see Section 5), we even found a considerable amount of instances where different publishing years were given for the same publication. The deviations can be due to mistakes in the references of a publication, but they can also be caused by an erroneous output of the GROBID tool or by errors in our programming scripts.

For validating the mapping, we generate a triplet representing a citation, consisting of the last name of the first author, the publication year and the title. Authors and titles are normalized by a conversion to lower case, the reduction to the letters [a-z] (thus deleting non-alphabetic characters and whitespace) and a limitation of the normalized title to a maximum length of 40 characters. For example, the triplet representing our present paper is "lindemann_2018_lexbibacorpusandbibliographyofmetalexico". The mapping is considered valid if one of several validity categories are fulfilled; if, for example, three triplets are found where the (non-normalized) Levenshtein distance of the titles is ≤ 8; the Levenshtein distance of the authors is ≤ 2, and the publishing years may differ by one year. Note that this restriction implies that documents with less than three citations are filtered out from the citation analysis.

5 Preliminary Experiments

For a study on the overlap of topics and citations between Digital Humanities (DH) and lexicography, an e-science corpus has been built and processed applying the methodology described in section 4. Table 3 shows the composition of the lexicography subcorpus, which is identical to the LexBib test set mentioned in section 2.1. The DH publications stem from four major DH journals and a DH handbook (see Lindemann, Kliche & Kutzner 2018 for the complete reference).

The results of the computational text analysis performed on that corpus confirm an initial hypothesis that despite a very small overlap of the citations (i.e. in spite of the fact that authors from DH and lexicography hardly cite each other), quite a wide range of overlapping topics and terms is found. Topic clustering disclosed a very significant amount of topics where publications from both disciplines are found among the publications with the highest weight for a topic; in other words: a list of topics that can be regarded as important for both DH and lexicography. We visualized the results of the topic modeling in the table-like model in Figure 2. Columns represent the topics and contain the top 100 most relevant publications for a topic. Publications from lexicography are highlighted in green; publications from DH in purple. The figure shows that for some topics the top 100 relevant publications belong (nearly) exclusively to one of the two domains, while in many other cases publications from both domains appear.

Table 3: Composition of the LexBib full text collection (test set).

Journals	***915***
Lexikos	376
International Journal of Lexicography	282
Dictionaries (Journal of the DSNA)	257
Conference Proceedings	***984***
Euralex	782
eLex	202
Handbooks	***157***
HSK 5/4 (Gouws et al. 2013)	110
Routledge Handbook (Fuertes Olivera 2018)	47
Total	***2,056***

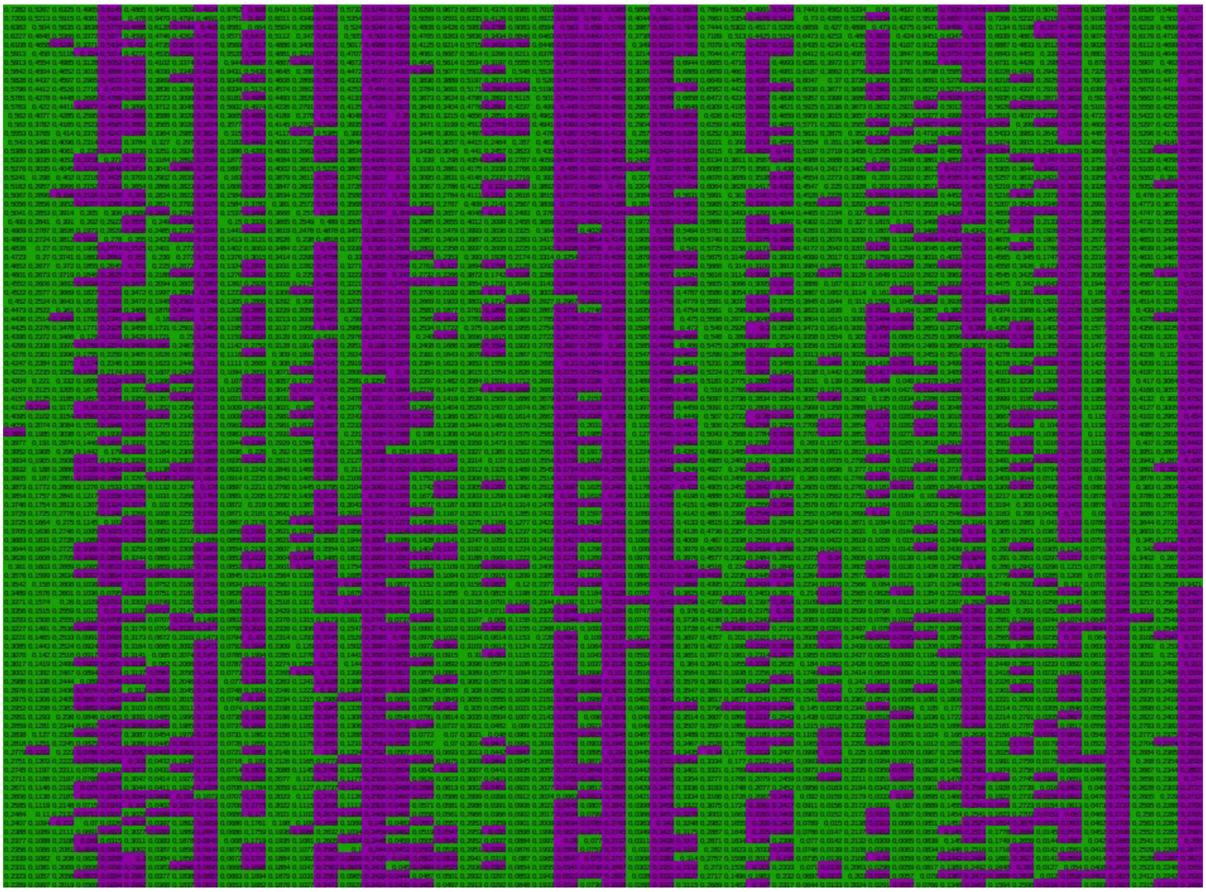

Figure 2: Visualization of topic clusters and their relevance in the DH/lexicography subcorpora.

In the following, we will focus on some details of the term extraction results. The term extraction tool produced a list of term candidates for each of the two subcorpora, Digital Humanities (DH), and Lexicography (Lexicog), ranked by their *termhood* in relation to the frequency in the reference corpus, the BNC. Table 4 lists the top 25 terms for DH, Lexicog, and their overlap, i.e. lexicography terms (out of the top 1,000) also found in the DH top 500, sorted by their termhood ratio in comparison to their frequency in the BNC.

Table 4: Term candidates extracted from the DH and Lexicography subcorpora.

Top DH Terms	*Top Lexicog Terms*	*Top LexTerms found in DH*
website	dictionary article	website
pdf	access structure	lemmatization
xml	dictionary user	wordnet
stemma	lemma sign	reference corpus
text mining	text reception	corpus query
authorship attribution	multiword	search engine
blog	website	internet
text classification	dictionary consultation	web site
cyberinfrastructure	word formation	web page
search engine	lexicography	text box
feature selection	corpus evidence	print version
url	text production	subcorpus
classification accuracy	lemmatization	frequency list
web page	dictionary research	crowdsourcing
web site	function theory	web interface
php	article stretch	corpus research
crowdsourcing	word sketch	language documentation
open-source	wordnet	word alignment
base text	reference corpus	hyperlink
internet	translation equivalent	Wikipedia
text categorization	dictionary information	search interface
test text	pdf	blog
text reuse	lemma list	source word
book history	dictionary making	text genre
metadata	definition	word sense disambiguation

After manually validating the 500 most salient DH terms, we performed the same term extraction procedure for every year of publication and measured the intersection of the Lexicog terms and the top 500 DH terms. As Figure 3 shows, the amount of DH terms (top 500) in the diachronically indexed Lexicog subcorpora (top 1,000 candidates for each year) shows an upward tendency. Three years appear to be the most salient ones in that respect, and these happen to be years when a conference of the eLex series has taken place.

In order to verify this observation, we had a closer look at the publications contained in the eLex conference proceedings: and indeed, the term extraction results show a higher representation of DH-relevant terms in the eLex subcorpus than in the Lexicog corpus in general (see Figure 4).

These trend analyses are only two examples of applications that imply a re-use of text analysis outcomes in the first place meant as additional publication metadata for an online bibliography; two examples of insights driven by quantitative text analysis that require a minimal effort once the lexicography e-science corpus is built and processed in the described way.

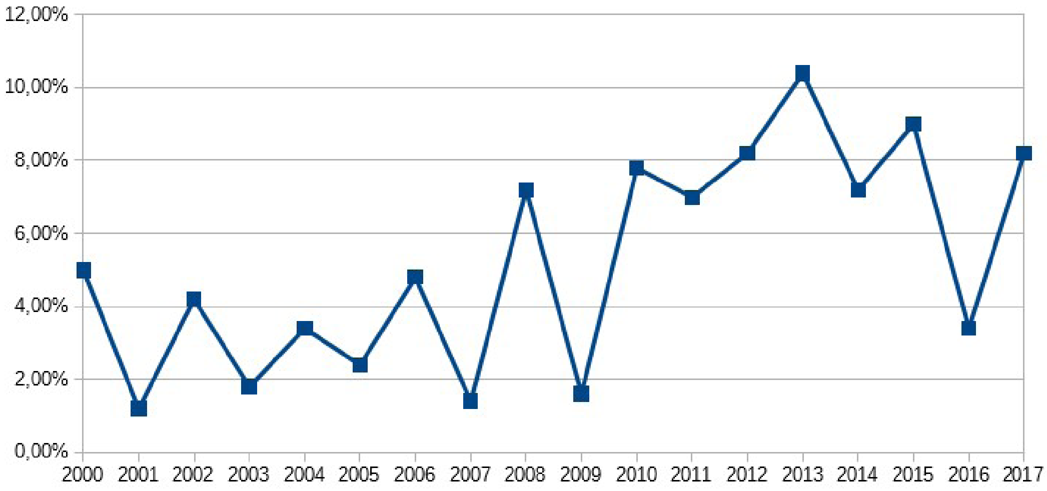

Figure 3: Overlap of term candidates from the DH and Lexicog subcorpora.

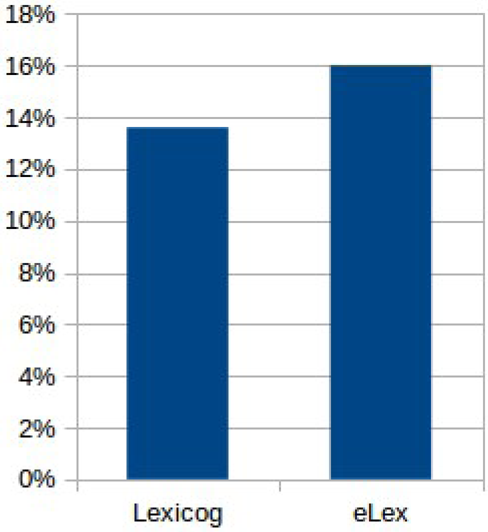

Figure 4: Overlap of DH and lexicography term candidates in the Lexicog vs. the eLex subcorpora.

6 Outlook

We think that lexicography is a discipline important enough as to deserve a well-structured and well-maintained bibliography as research infrastructure, and, at the same time, that it is a discipline small enough as to allow a collective reflection and a continuous evaluation of a project of this kind. The main idea is that LexBib should become a collaboratively run and widely used resource. We will call to the community for collaboration, e.g. regarding author grounding, i.e. ORCID indexing, and completion of author pages, and the evaluation of automatic keyword indexation and automatic summaries. In case the LexBib user community reaches a critical mass for introducing user generated content, we will also study the possibility of enabling user comments or discussion threads on LexBib items.

References

Ahmad, K., Davies, A., Fulford, H & Rogers, M. (1992). What is a term? The semi-automatic extraction of terms from text. In *Translation Studies - An Interdiscipline. Selected papers from the Translation Studies Congress, Vienna*, 267 – 278.

Ahumada, I. (2006). *Diccionario bibliográfico de la metalexicografía del español. Vol. I: orígenes-año 2000*. Jaén: Universidad de Jaén.

Ahumada, I. (2009). *Diccionario bibliográfico de la metalexicografía del español. Vol II: años 2001-2005*. Jaén: Universidad de Jaén.

Ahumada, I. (2017). *Diccionario bibliográfico de la metalexicografía del español. Vol III: años 2006-2010*. Jaén: Universidad de Jaén.

Ahumada, I. (2016). Metalexicografía del español: clasificación orgánica y tipología de los diccionarios en el Diccionario Bibliográfico de la Metalexicografía del Español (DBME). In *Anuario de estudios filológicos*, (39), 5–24.

Burch, M., Pompe, D., & Weiskopf, D. (2015). An analysis and visualization tool for DBLP data. In *Proceedings of the 19th International Conference on Information Visualisation (iV),* IEEE, 163–170.

De Schryver, G.-M. & R. Lew (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicography*, (27,4), 341-359.

De Schryver, G.-M. (2012). Trends in Twenty-Five Years of Academic Lexicography. *International Journal of Lexicography*, (25,4), 464–506.

De Schryver, G.-M. (2009). Bibliometrics in Lexicography, *International Journal of Lexicography*, (22,4), 423–465.

Fuertes-Olivera, P. A. (Ed.). (2018). *The Routledge Handbook of Lexicography*. Routledge Handbooks in Linguistics. London: Routledge.

Gouws, R. H., Heid, U., Schweickard, W., & Wiegand, H. E. (Eds.). (2013). *Dictionaries. An International Encyclopedia of Lexicography*. HSK 5/4. Berlin, Boston: De Gruyter Mouton.

Heid, U. (2014). Natural Language Processing techniques for improved user-friendliness of electronic dictionaries. In A. Abel, C. Vettori, N. Ralli (Eds.), *Proceedings of the 16th EURALEX International Congress*. Bolzano: EURAC research, 47-61

Jacinto García, E. J. (2016). El Diccionario Bibliográfico de la Metalexicografía del Español como obra de consulta: estructura, fuentes y funciones. *Anuario de estudios filológicos*, (39), 147–169.

Ley, M. (2002). The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *String Processing and Information Retrieval*, Lecture Notes in Computer Science. Presented at the International Symposium on String Processing and Information Retrieval, Berlin, Heidelberg: Springer, 1-10.

Lindemann, D., Kliche, F. & Kutzner, K. (2018). Lexikographie: Explizite und implizite Verortung in den Digital Humanities. In G. Vogeler (Ed.), *5. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. DHd 2018 - Kritik der Digitalen Vernunft, Konferenzabstracts*. Köln: Universität Köln, 257-261.

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. Amherst, MA: University of Massachussetts. Retrieved from http://mallet.cs.umass.edu/

Möhrs, C. (2016). Online Bibliography of Electronic Lexicography. The Project OBELEXmeta. In T. Margalitadze & G. Meladze (Eds.), In *Proceedings of the 17th EURALEX International Congress: Lexicography and Linguistic Diversity*. Presented at the XVII International Euralex Conference, Tbilisi: Tbilisi State University, 906-909.

Porta Zamorano, J. (2016). DBME_3: Adquisición de datos, composición y base de datos Nebrija-Valdés. *Anuario de estudios filológicos*, (39), 349–355.

Rösiger, I., Bettinger, J., Schäfer, J., Dorna, M., & Heid, U. (2016). Acquisition of semantic relations between terms: how far can we get with standard NLP tools? In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, 41-51

Rösiger, I., Schäfer, J., George, T., Tannert, S., Heid, U., & Dorna, M. (2015). Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries. In I. Kosem, M. Jakubíček, J. Kallas, & S. Krek (Eds.), *Proceedings of the eLex 2015 conference*. Herstmonceux Castle, United Kingdom, Ljubljana; Brighton: Trojina, Institute for Applied Slovene Studies; Lexical Computing Ltd.

Romary, L. & Lopez, P. (2015). GROBID – Information Extraction from Scientific Publications. ERCIM News, Scientific Data Sharing and Re-use, (100).

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester.

Theofilidis, A. (2018). Methoden der inhaltlichen Erschließung neuer Fachdomänen auf der Grundlage von Termextraktionsverfahren. In P. Drewer, F. Mayer, K.-D. Schmitz (Eds.), *Terminologie und Texte. Akten des DTT-Symposion 2018, Mannheim*. München, Karlsruhe, Köln: Deutscher Terminologie-Tag e.V., 87-97

Weber, A., Reuther, P., Walter, B., Ley, M., & Klink, S. (2006). Multi-Layered Browsing and Visualisation for Digital Libraries. In *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 520-523

Wiegand, H. E. (2006a). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 1: A-H*. Berlin, Boston: De Gruyter.

Wiegand, H. E. (2006b). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 2: I-R*. Berlin, Boston: De Gruyter.

Wiegand, H. E. (2007). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 3: S-Z*. Berlin, Boston: De Gruyter.

Wiegand, H. E. (2014). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 4: Nachträge*. Berlin, Boston: De Gruyter.

Wiegand, H. E. (2015). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 5: Register*. Berlin, Boston: De Gruyter.

Wiegand, H. E., Beißwenger, M., Gouws, R. H., Kammerer, M., Mann, M., Storrer, A., & Wolski, W. (Eds.). (2017). *Wörterbuch zur Lexikographie und Wörterbuchforschung, Band 2*. Boston: Walter de Gruyter.

Wiegand, H. E., Beißwenger, M., Gouws, R. H., Kammerer, M., Storrer, A., & Wolski, W. (Eds.). (2010). *Wörterbuch zur Lexikographie und Wörterbuchforschung, Band 1*. Berlin: De Gruyter.

Zeni, N., Kiyavitskaya, N., Mich, L., Mylopoulos, J., & Cordy, J. R. (2007). A Lightweight Approach to Semantic Annotation of Research Papers. In *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science.. Berlin, Heidelberg: Springer, 61-72.