What's in A Name? The Case of *Albanisch-Albanesisch* and Broader Implications

Erhard Hinrichs
Seminar für Sprachwissenschaft
Eberhard Karls Universität
Tübingen, Germany
erhard.hinrichs@uni-tuebingen.de

Alex Erdmann
Department of Linguistics
The Ohio State University
Columbus, Ohio, USA
erdmann.6@buckeyemail.osu.edu

Brian Joseph
Department of Linguistics
The Ohio State University
Columbus, Ohio, USA
joseph.1@dosu.edu

Abstract

This paper offers a use case of the CLARIN research infrastructure (Hinrichs and Krauwer, 2014) from the fields of historical linguistics and the history of linguistics. Using large electronically available corpora of historical English and German, it investigates differences in terminology used in the two languages when referring to the people and the language of Albania. The search and data exploration tools that are available for the DTA and the DWDS corpora as part of the CLARIN-D infrastructure (Hinrichs and Trippel, in press) make it possible to determine semantic change for the terminology under consideration. The paper concludes with a discussion of broader implication of the present use case for the use of historical corpora and the functionality of query tools needed for digital humanities research.

1 Introduction

The name for the country that lies on the western coast of the central part of the Balkan peninsula in south-eastern Europe, as well as for its people and its language, presents interesting variation in both German and, to a far lesser extent, English, raising questions about the nature and the chronology of the variation. The country in question is, in its usual form today in English, *Albania*, the people *Albanians*, and the language *Albanian*, and on the German side, the most usual terms nowadays are *Albanien*, *Albaner*, and *Albanisch*. However, if one looks at materials from a century ago, the picture is somewhat different in that variant forms of the substantival stem are rather widespread in German: *Albanese-* and *Albanier-* for the people and *Albanisch-* and *Albanesisch-* for the language. Even in English, in one author, linguist Leonard Bloomfield (Bloomfield (1914; 1933), the variant *Albanese* for the language name is encountered, as in the following quote from Bloomfield (1933, p. 14), with for emphasis *Albanese* added by the authors:

> *In the same way, finding all these languages and groups (Sanskrit, Iranian, Armenian, Greek,* ***Albanese****, Latin, Celtic, Germanic, Baltic, Slavic) resemble each other beyond the possibility of mere chance, we call them the Indo-European family of languages.*

Since Bloomfield's first academic mentor in linguistics was the Austrian-born Indo-Europeanist Eduard Prokosch and since Bloomfield spent part of his postdoctoral training with leading Indo-European scholars at the University of Leipzig and at the University of Göttingen in 1913-14, one cannot help but wonder whether Bloomfield's choice of the term *Albanese* in place of *Albanian*, the term used by other

Erhard Hinrichs, Alex Erdmann and Brian Joseph 2017. What's in a name? The case of albanisch-albanesisch and broader implications. *Selected papers from the CLARIN Annual Conference 2016.* Linköping Electronic Conference Proceedings 136: 38–53.

contemporary English-speaking scholars, has its roots in the German scholarly tradition. The hypothesis that Bloomfield borrowed the term *Albanese* from German scholarly tradition presupposes that the lemma *Albanese* was, in fact, the preferred way to refer to people of Albania in German at the beginning of the 20th century. This in turn raises the question about the usage patterns in German of the nouns *Albaner-* versus *Albanese-* and the related adjectival forms of *Albanisch-* and *Albanesisch-* at that time.

With the increased availability of large electronic historical language corpora, it has become significantly easier to trace the usage patterns of words and to document changes in word meaning over time. In the present paper, three electronic collections of historical and contemporary German will be consulted to answer these questions and to shed some light on the variation noted above in both German and English: the Google Books collection of digitized English and German books (henceforth: GBCE and GBCG, respectively), the Deutsche Text Archiv (henceforth: DTA; www.dta.de) (Geyken et al., 2011), and the corpus of the Digitales Wörterbuch der deutschen Sprache (www.dwds.de) (Geyken, 2007), both available at the CLARIN Center at the Berlin-Brandenburg Academy of Sciences (BBAW) as part of the CLARIN-D research infrastructure.

The remainder of this contribution is structured as follows: Section 2 contrasts the usage of the term *Albanese* in English and German by consulting the Corpus of Historical American English (COHA) (Davis, 2012), the Google books collections for English and German (Michel et al., 2012) and DTA collections for German. Section 3 utilizes the DiaCollo tool (Jurish 2015) to trace changes in meaning over time for the German words under consideration. Sections 4–6 discuss some methodological issues and broader implications of the present use case and summarize the results.

2 Comparative Study of Historical Corpora for English and German

A comparative diachronic study of the terms *Albania* versus *Albanien*, *Albanian*/*Albanese* versus *Albanisch*/*Albanesisch*, and *Albanians* versus *Albaner*/*Albanesen* needs to consult historical corpora of both English and German. Since the focus is on American English, the COHA corpus of Historical American English is the most relevant English data source for the present investigation. For historical German, the DTA corpus collections with texts ranging from 1600 to 2000 is used as a data source. Both the COHA and the DTA corpus collections are linguistically annotated and include lemma and part-of-speech information. In addition to the two linguistically annotated corpora, the Google Books collections for English and German were consulted with the help of the Google Ngram viewer. The inclusion of these collections as data sources is motivated by the size of the Google Books collections.

2.1 Results for the COHA corpus of Historical American English

The COHA corpus is a balanced corpus of 400 million words with texts ranging from 1810 to 2000. It is currently the largest corpus of its kind and contains texts from the following genres: fiction, academic writing, magazines and newspapers. For the search string *albanian*[1] , COHA returns 387 occurrences in total, with 10 data points for the 19th century. The query term *albanese* yields a total of 28 occurrences for the following decades (with frequencies shown in parentheses: 1830(1), 1880 (1) 1940 (12), 1950 (4), 1960 (1), 1970 (3), 1980 (5), and 1990 (1). Examination of the linguistic context for each occurrence reveals that only the two data points from the 19th century refer to a person from the country of Albania. All other data points refer to someone named Albanese. These findings show that mere frequency counts can be quite misleading and need to be followed up with an inspection of the context of use for each occurrence or require high-quality named-entity tagging that would identify the proper name usage of the search term.

2.2 Results for the Google Books Collection for English and German

Figure 1 presents the results for all word forms of *Albanian-* and *Albanese-* for the GBCE corpus of English and confirms, as expected, that the former outranks the latter by a wide margin for entire period covered by the GBCE. As is the case for the COHA corpus, almost all data points for *Albanese* in the GBCE concern persons with the last name *Albanese*, rather than persons from Albania.

[1]The search terms for *albanian* and for *albanese* need to be submitted in all lowercase letters in the COHA interface.

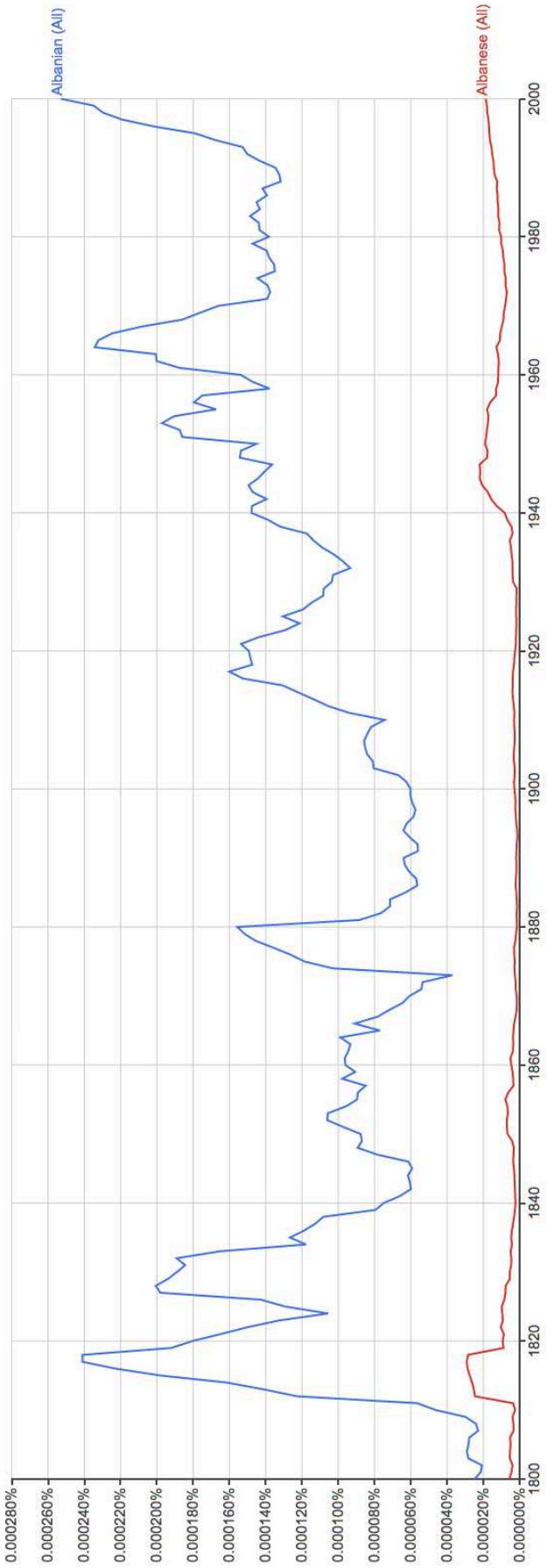

Figure 1: GBCE search results for *Albanian*/*Albanese*.

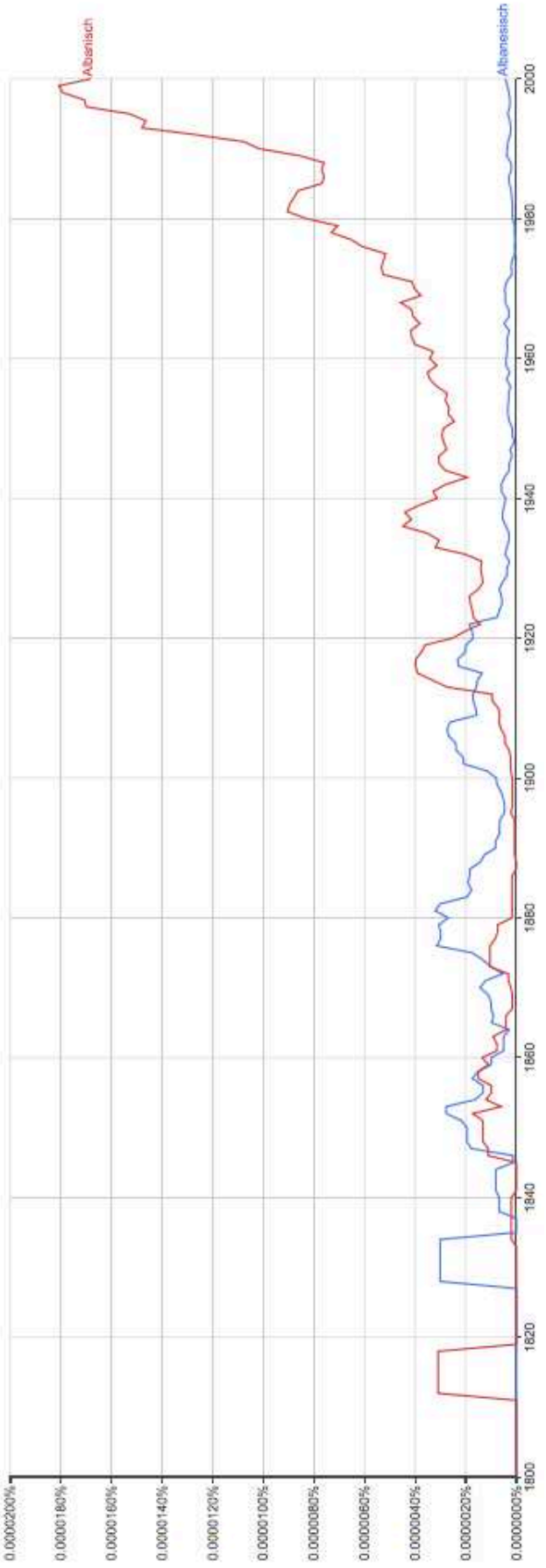

Figure 2: GBCG search results for *Albanisch*/*Albanesisch*.

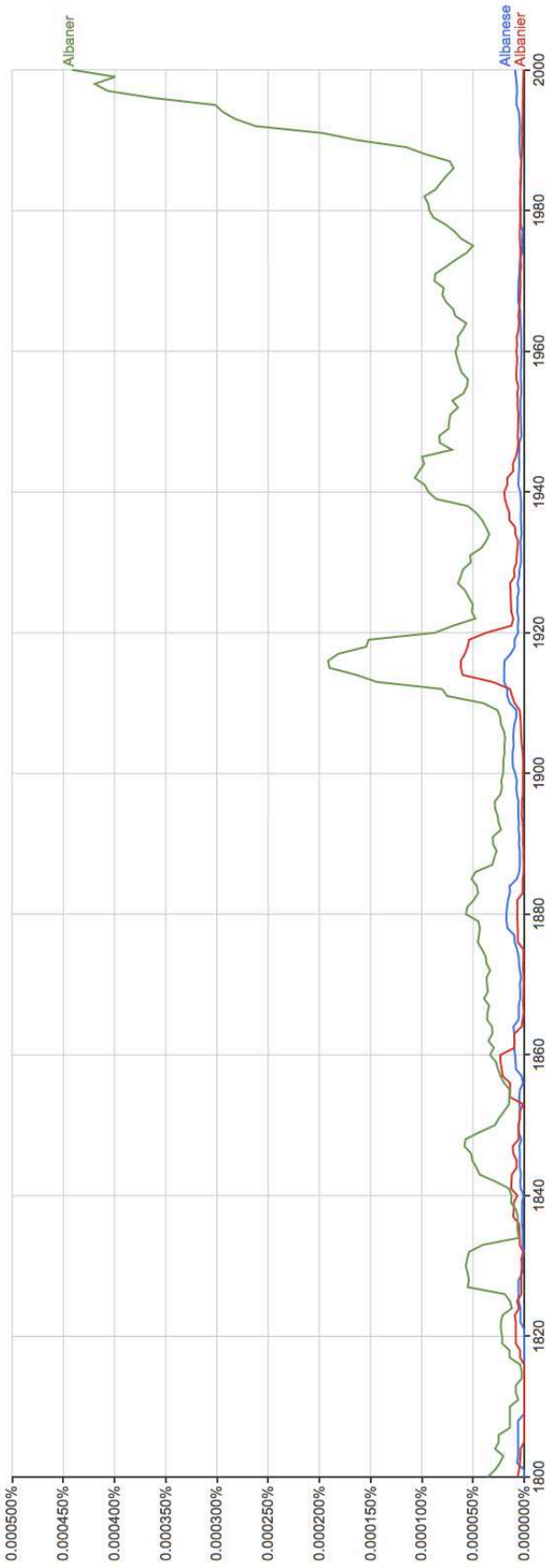

Figure 3: GBCG search results for *Albaner*/*Albanier*/*Albanese*.

Lemma	Frequency Count	Earliest DP	Latest DP
Albanien	109	1627	1913
Albanisch	97	1650	1913
Albanesisch	19	1789	1913
Albaner	61	1663	1913
Albanier	21	1661	1881
Albanese	36	1789	1913

Table 1: DTA query results.

Figure 2 shows the results for all word forms of *Albanisch-* and *Albanesisch-* for the GBCG. *Albanesisch-* outranks *Albanisch-* in relative frequency for most of the 19th century and up until 1914 and then shows a steady decline for the remainder of the century. This result increases the likelihood that Bloomfield may have adopted this term from his German-speaking academic teachers and during his postdoctoral stay in Germany in 1913/14. However, the search results for the nouns *Albaner-* and *Albanese-* in Figure 3 differ from the results in Figure 2 in that the former outranks the latter for entire period covered by the GBCG.

Are we to conclude from Figure 3 that *Albaner* was the preferred term of reference for persons from Albania, with *Albanese* and *Albanier* secondary variants? The mere frequency counts in Figure 3 do not suffice to give a reliable answer to this question. Rather, close inspection of the linguistic contexts for each occurrence of the terms in question is required to determine the intended referent. While the Google Books Ngram Viewer provides links to the digitized objects for each occurrence found for the search terms under consideration, there are a number of limitations due in part to Google's proprietary page ranking algorithm and in part to copyright restrictions. Copyright restrictions prohibit easy and complete inspection of the underlying digitized texts since for some sources only metadata can be provided. Presentation of the data via page rank, rather than by chronological order, makes it difficult to easily detect systematic changes in word meaning for the search terms in question.

2.3 Results for the DTA Corpus

The DTA contains German texts ranging from 1610 to 1900. The texts have been digitized and transliterated, using a high-precision double-keying method. The archive is still under construction. The version used for the present study dates from September 2016 and consists of 142,348,468 lexical tokens with 993,828,135 Unicode characters that are taken from 595,929 digitized pages and 2,448 different published works. The texts represent different genres, including novels and other literary works, scientific and journalistic texts.

The DTA corpus does not suffer from the same limitations as the GBCG. Search results can be rendered in ascending or descending chronological order with open access to all digitized texts via a web application supporting any web browser; seamless linking of facsimiles, digitized object data with the search term highlighted in red in its surrounding context, as well as complete and high-quality metadata all support a comprehensive and reliable inspection of the entire data set. Table 1 provides the frequency counts for the same set of words investigated in the GBCG corpus.

Inspection of the linguistic contexts for all DTA data points reveals that the adjectival and nominal uses of [*a*|*A*]*lbanesisch-* refer to the country or the language spoken in Albania, and all instances of *Albanese* and *Albanier* refer to persons from Albania. By contrast, all instances of the lemma [*a*|*A*]*lbaner-* in the DTA refer to people or locations north of Rome and not to people from Albania, which is the present usage of this lemma. Typical bigrams found in the DTA include *Albaner See* ('Alban lake'), *Albaner Gebirge* ('Alban mountains'), *Albaner Könige* ('Alban kings') as local rivals of the Roman Empire.

Unlike the other three terms, the lemma *Albanisch* has two distinct senses in the DTA, with some of its uses referring to entities related to the territory north of Rome and other instances referring to entities

related to Albania. Examples (1) and (2) illustrate these distinct uses, with the term *albanische* in (1) referring to the location north of Rome and in (2) referring to the language spoken in Albania.

(1) *Dass Alba als Haupt- und Muttergemeinde galt, ist gewiss, und bloss in diesem Sinn*
That Alba as main and mother community featured is certain and only in this sense
wird Rom auch als albanische Colonie bezeichnet.
is Rom also Alban colony considered

'That Alba featured as main and mother community is certain, and only in this sense is Rome considered an Alban colony.

source: Mommsen (1854), p. 29

(2) *Die albanische Sprache ist der älteste griechisch aeolische Dialect.*
the Albanian language is the oldest Greek aeolic dialect

'The Albanian language is the oldest Greek aeolic dialect.'

source: Libelt (1828), p. 430

The use of the term *Albanesisch* to refer to the language spoken in Albania may therefore at least partly be motivated by the well-attested and well-motivated pragmatic strategy of trying to avoid ambiguity. Such an ambiguity would have arisen if the term *Albanisch* would have been used instead. The fact that the bigram *albanesische Sprache* occurs with higher frequency than the *albanische Sprache* in the GBCG corpus, as shown in Figure 4, provides cross-corpus evidence for this strategy being at work. Notice also that the bigram *albanesische Sprache* maintains higher frequency in the GBCG corpus for about the same time period during which the unigram *albanesisch* outnumbers the *albanisch*.

Likewise, the choice of the terms *Albanese-*, and *Albanier-* to refer to the people from *Albania* and of the term *Albaner-* to refer to people from a region north of Rome suitably avoids ambiguity of reference.

For the period covered by the DTA corpus, the language of Albania was referred to as *Albanesisch* and the people were referred to as *Albanesen* or *Albanier*. These corpus findings support the hypothesis that Bloomfield's use of the English term *Albanese* may be due to his close contacts with German scholars who would have used the German cognate. What still remains to be accounted for is the linguistic change that took place in the 20th century, when the term *Albanesisch* was replaced by *Albanisch* as the name of the language, and the term *Albaner* replaced *Albanesen* and *Albanier* as the name of the people of Albania. The DWDS corpus and the tools DiaCollo (Jurish, 2015), which are made available by the CLARIN center at the BBAW make it possible to trace these two changes.

3 Tracking Semantic Change in the DTA and DWDS Collections

The web application DiaCollo collects for a given query term sets of collocates for regular time slices within a text collection. Changes in collocation behavior of a target word are one diagnostic for detecting changes in word meaning over time since the choice of collocates help to disambiguate the meaning of a word.

Figure 5 contrasts the collocates found by DiaCollo for the word *albanisch* in the DTA and DWDS corpus collections. DiaCollo supports continuous word cloud animations for the entire time interval chosen for a particular query. For the query at hand, the time interval is specified by the parameter-value setting `DATE(S):1670-2010`. A continuous DiaCollo animation over the entire interval is available at URL `http://kaskade.dwds.de/dstar/public/diacollo/`. In this paper, we can only show individual frames from this more complete animation. The upper panel in Figure 5 shows the noun *Erz* ('ore') as the only collocate in the DTA texts for the decade (`SLICE:10`) starting with 1890; the lower panel provides the five strongest collocates (`KBEST:5`) for the term *albanisch* for the decade of 1960-1970. The collocates are shown as word clouds, which are one of the visualization options offered by the DiaCollo tool and chosen for this query by the parameter setting `FORMAT:cloud`. In the queries shown in Figure 5, the collocates are grouped by lemmas (rather than word forms) and filtered by the part-of speech label `NN` (short for: *normal noun*). This label is part of the Stuttgart-Tübingen tagset (STTS; Schiller et al. (1995)) that has been used for morph-syntactic tagging of the DTA and DWDS corpus collections.

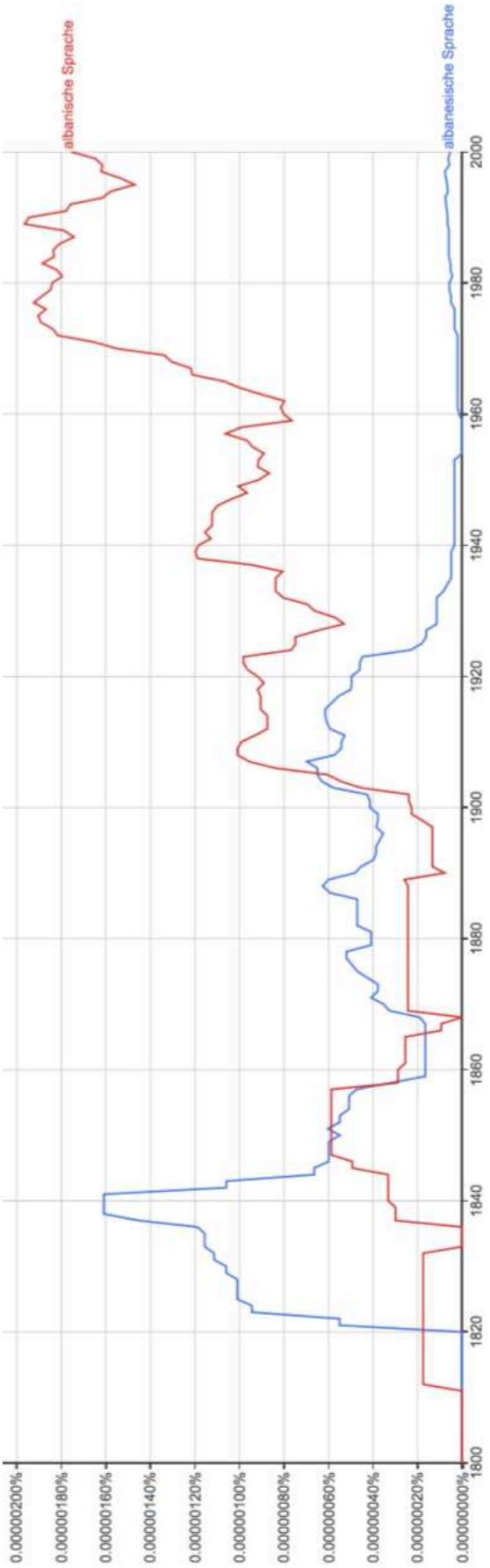

Figure 4: Bigram Comparison of *albanische Sprache* und *albanesische Sprache* in the GBCG.

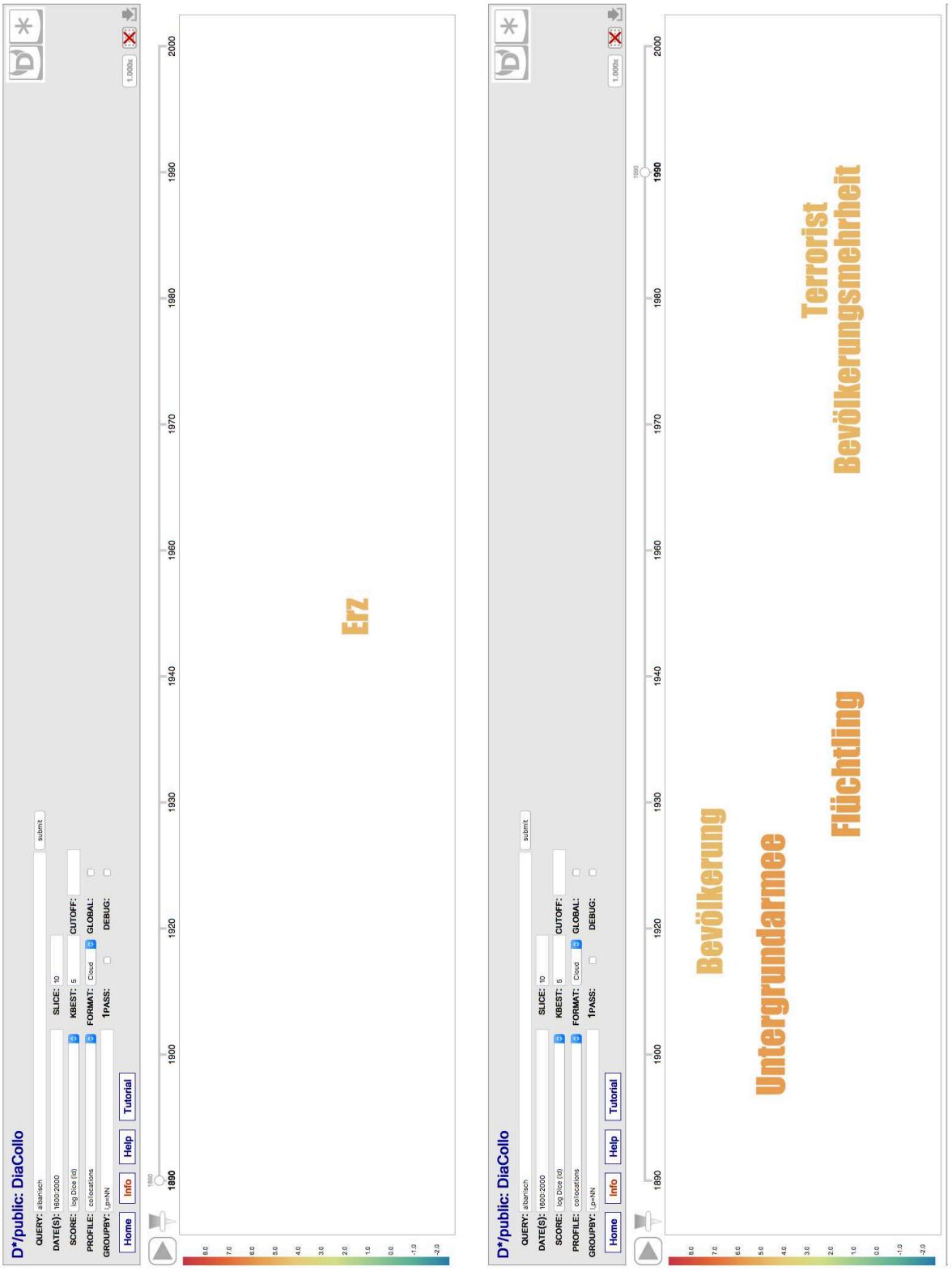

Figure 5: Collocations of *albanisch* in the DTA and the DWDS corpora: decades of 1890 (left panel) versus 1960 (right panel).

Collocate Noun	From	To
Erz ‘ore’	1890	1910
Patriarchat ‘patriarchy’	1910	1920
Regierung ‘government’	1910	1920
Nationalversammlung ‘national assembly’	1920	1930
Aufständische ‘rebels’	1940	1960
Telegrafenagentur ‘telgraph agency’	1940	1960
Front ‘battle line’	1940	1960
Regierung ‘government’	1940	1960
Grenze ‘border’	1940	1960
Kp Führer ‘Communist Party leader’	1950	1970
Parteiführer ‘Party leader’	1950	1970
Stalinist ‘Stalinist’	1950	1970
Spalter ‘divider’	1950	1970
Marxismus-Leninismus	1950	1970
Ausschluß ‘exclusion’	1970	1980
Partei ‘party’	1970	1980
Volk ‘people’	1970	1980
Antrag ‘petition’	1970	1980
Serbe ‘Serbian’	1980	1990
Jahr ‘year’	1980	1990
Flüchtling ‘refugee’	1990	2000
Bevölkerung ‘population’	1990	2000
Bevölkerungsmehrheit ‘population majority’	1990	2000
Untergrundarmee ‘underground army’	1990	2000
Terrorist ‘terrorist’	1990	2000

Table 2: DTA and DWDS query results for NE collocates of *albanisch*.

Collocation strength is measured by a suite of statistical scores that includes the scaled log-Dice coefficient (`SCORE:log Dice (ld)`). The scaled log-Dice score is defined by the following equation, due to Rychlý (2008), where f_1,f_2, and f_{12} present the raw frequency counts of the collocate, the query term, and of the joint occurrences of the query term and the collocate, respectively.[2]

$$score_{ld} = 14 + log_2(\frac{2 * (f_{12} + \epsilon)}{((f_1 + \epsilon) + (f_2 + \epsilon))} \quad (3)$$

The color spectrum, shown to the left of the word cloud panels in Figure 5, is correlated with the log Dice scores from -2 to 10, in ascending order of collocation strength. Hence, the color used to display a given collocate in a DiaCollo word cloud indicates the collocation strength of the word: *Flüchtling* ‘refugee’ and Untergrundarmee ‘underground army’ are, therefore, the strongest collocate for *albanisch* for the decade 1990-2000 shown in the lower panel in Figure 5.

While the DiaCollo search spans over the time frame of 1610 - 2000, so as to include the time coverage of the DTA and the DWDS, the first decade that yields a common noun with sufficient collocation strength for *albanisch* is the one beginning with 1890. This could either mean that prior to 1890 the lemma *albanisch* did not occur with sufficient frequency itself or that the set of co-occurring lemata was too widely dispersed. Table 2 lists the set of common noun (NN) collocates identified by DiaCollo for the time period from 1890 to 2000 and records the decade during the first and last occurrence for each

[2]Other score functions available in Diacollo include pointwise mutual information and binomial log-likelihood ratio.

collocate. Inspection of the linguistic contexts, in which *albanisch* co-occurs with the noun *Erz* ('ore') reveal that the ore referred to comes from the Alban region north of Rome. By contrast, for all collocates listed for the 20th century in Table 2, *albanisch* is linked to the country of Albania. This suggests that the change in meaning originated at the turn of the 19th and 20th century. The frequent change in the five strongest collocates per decade is indicative of the many changes in the history of Albania during 20th century.

The second semantic change involving the term *Albaner* can also be traced with the help DiaCollo tool. The upper panel in Figure 6 shows the noun *Römer* ('Romans') as the only collocate in the DTA texts for the decade (`SLICE:10`) starting with 1670; the lower panel provides the five strongest collocates (`KBEST:5`) for the term *Albaner* for the decade of 1990-2000.

The disjoint sets of collocates between the decades starting with 1670 and 1990 indicate that the meaning of the term *Albaner* has shifted from referring to people or other entities associated with a territory north of Rome, to the people or other entities from the Balkan country Albania, with the collocates *Serbe* ('Serb'), *Kfor*, *Kfor-Soldat* ('Kfor-soldier'), *Provinz* ('province'), and *Vertreibung* ('forced migration') all salient lemmas at that time, due to the Balkan wars.

Collocate Noun	**From**	**To**
Römer 'Roman'	1670	1880
Gebirge 'mountain range'	1700	1970
Stein 'stone'	1700	1970
Berg 'mountain'	1910	1990
Jahr 'year'	1910	1990
Serbe 'Serb'	1910	2000
Provinz 'province'	1990	2000
Vertreibung 'forced migration'	1980	2000
Kfor 'Kfor'	1980	2000
Kfor-Soldat 'Kfor soldier'	1980	2000
Friedenstruppe 'peace keeping force'	1990	2000

Table 3: DTA and DWDS query results for NN collocates.

Collocate Noun	**From**	**To**
Rocca 'Rocca'	1850	1970
Jugoslawien 'Jugoslavia'	1860	1980
Kosovo 'Kosovo'	1910	2000
Mazedonien 'Macedonia'	1980	1990
Pristina 'Pristina'	1980	1990
Rugova 'Rugova'	1980	1990
UCK 'UCK'	1980	1990
Kosovska Mitrova 'Kosovska Mitrova'	1990	2000
Serbien 'Serbia'	1990	2000

Table 4: DTA and DWDS query results for NE collocates.

Table 3 lists the set of common noun (NN) collocates identified by DiaCollo for the time period from 1670 to 1990 and records the decade during of first and last occurrence for each collocate. While most collocates are clearly indicative of a particular reading of the term *Albaner*, the collocates *Gebirge*, *Stein*, *Berg*, and *Jahr* are not. Examination of the linguistic contexts of the collocates reveals that with the exception of *Jahr*, where *Albaner-* refers to persons from Albania, for all other collocate nouns the query term refers to the Italian region north of Rome.

The term *Serbe* is the first collocate in chronological order that clearly shows the shift in meaning for

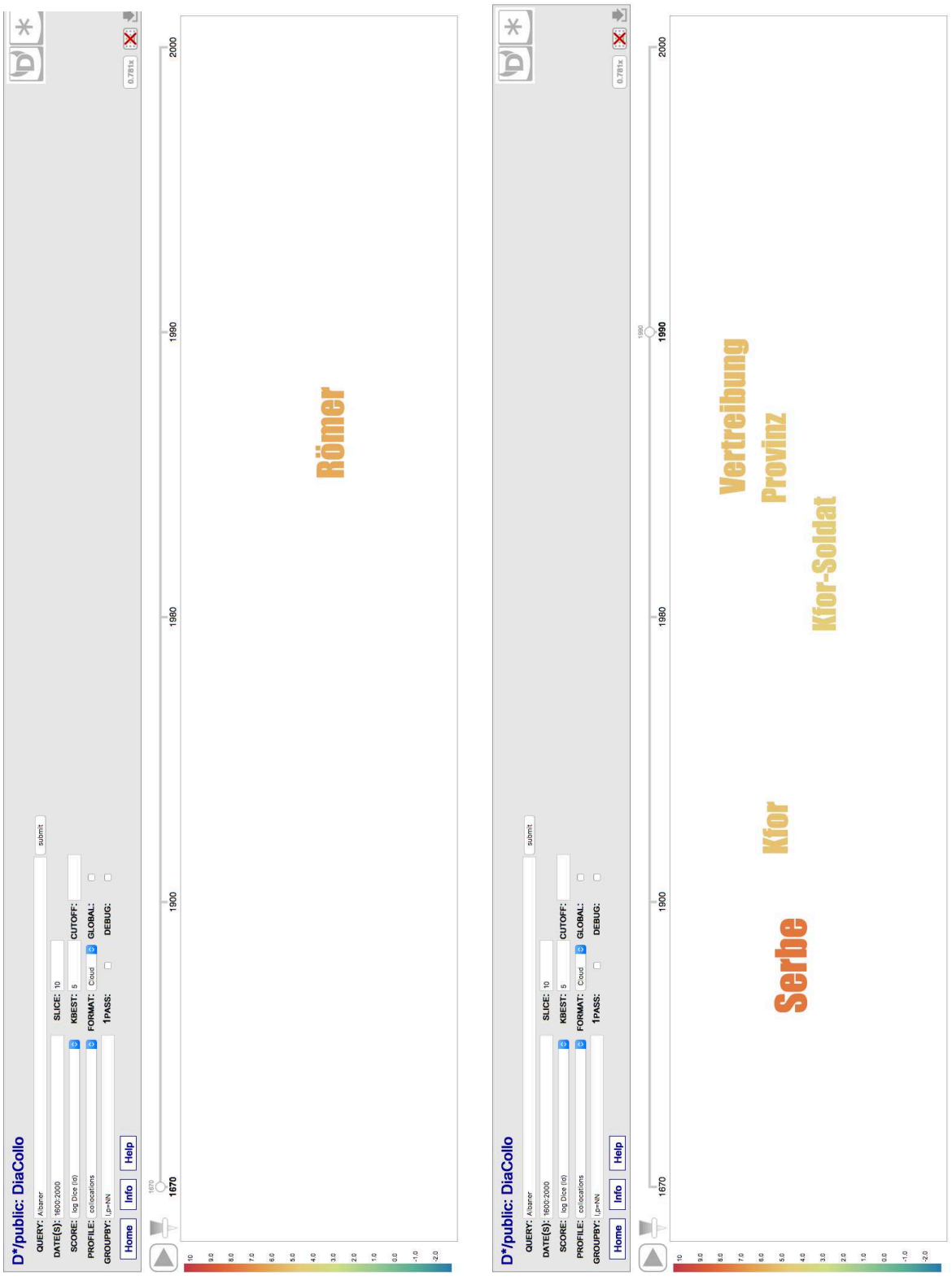

Figure 6: Collocations of *Albaner* in the DTA and the DWDS corpora: decades of 1670 (left panel) versus 1990 (right panel).

the term *Albaner*. Additional data points can be obtained by a DiaCollo query that filters for collocates belonging to part-of-speech category NE (short for: proper names in the STTS tagset). The results of this query are shown in Table 4. Among the NE collocates, the first occurrence of the collocate Jugoslawien for the query term *Albaner* provides evidence that the shift in meaning is starting already in the second part of the 19th century.

Data mining of the DWDS corpus of the 20th century provides additional evidence of a transitional period between the uses of the term *Albanese* at the beginning and of the term *Albaner* at the end of the 20th century. The DWDS contains a total of 33 occurrences of the term *Albanier* between 1914 and 1991. We suspect as well that homonomy avoidance, documented as a driving force in some semantic change (Hock and Joseph 1996: 224), may have been at work here.

4 Methodological Issues and Wider Implications

This corpus study was prompted by the writings of the American linguist Leonard Bloomfield and aimed at answering a specific research question concerning Bloomfield's use of the English term *Albanese*. The findings of the study have allowed us to track changes in the German lexicon for the language and the people of Albania. While these results are valuable in their own right, it is worthwhile to reflect on some lessons learnt in the course of this investigation. We will concentrate on those aspects and methodological issues which go beyond this particular use case and are applicable more widely to corpus studies based on diachronic language data.

In the present study, we consulted different corpora of various sizes and with varying degrees of linguistic post-processing, including spelling normalization, lemmatization, part-of-speech tagging, as well as collocation and bigram analysis. With Google's Ngram Viewer and DiaCollo word clouds we also utilized two kinds of visualizations to highlight relevant patterns in the data of interest. Such visualizations are essential, given the size of the Google Books and DTA corpus collections, and their utility extends well beyond the type of linguistic study that we are engaged in here to other areas of digital humanities research. In fact, they have given rise to a text-mining approach in its own right under the name of *culturomics*[3] and have have been widely applied in recent years to detect social dynamics of various kinds and different humanities disciplines.

As the proponents of culturomic methods have pointed out themselves, it is important to be aware of the limitations of the Google Books corpus data[4]. These limitations include errors due to optical character recognition (OCR), metadata quality, the opportunistic data collection method, and lack of lemmatization for the Google Books corpus for German. Historical texts are particularly prone to OCR errors, and metadata quality can be unsystematic if the texts included in the corpus are not first editions. Since there are no published evaluation results on these two issues for the Google Books corpus, its reliability is unknown. For the DTA, published results on both matters are available. (Haaf et al., 2013) overall accuracy rate of 99.9909% for a balanced DTA text sample, which implies on average 91 transcription errors in one million characters. The data collection policy for the DTA adheres to the principle that only first editions of individual texts are included in the DTA collection so as to ensure highly reliable metadata. The reliability of metadata is particularly important for search terms that occur with low frequency in a given corpus, as is the case for the set of lexical items under investigation in this paper. Unreliable metadata, due to automatic harvesting methods and/or lack of information about first editions, can lead to rather distorted results about historical trends in word usage and frequency.

Another difference between the Google Books collection for German and the DTA corpus collections concerns the lack of linguistic analysis and annotation for the former. The Google Books corpus for German is not lemmatized, and Google Ngram queries do not support regular expressions. Taken together, this means that only word form frequencies and no lemma frequencies can be displayed in the Google Ngram Viewer. Such limitations do not apply to DTA queries since the DTA data are lemmatized and the DTA's query language DDC (Jurish et al., 2014) supports searching for word forms (tokens) and lemmata. For morphologically rich languages like German this functionality is essential.

[3] See, inter alia, (Michel et al., 2012; Lieberman et al., 2007).

[4] See `//www.culturomics.org/Resources/faq` for a more in-depth discussion.

The search functionalities for the Google Books collection for German and for the DTA corpus differ not only in expressivity of the underlying query languages, but also in terms of the information that these two well-designed web applications convey. The main goal of the Google Ngram viewer is to visualize changes in the frequencies of ngrams over time. This is what makes it so attractive for diachronic corpus studies. The main focus of the DTA query interface, on the other hand, is on the seamless rendering and browsing of different textual views: a keyword-in-context view for a particular query along with a pointer to the relevant section of the underlying manuscript and its transcription. The keyword in context presentation of each data point is essential for a careful examination of the meaning of the query term and avoids the danger that arises if only unigram frequencies can be compared. In the course of the present investigation, we encountered precisely this type of situation, when the query term *Albanese* was a proper name, rather than the type of referent we were interested in. This potential error was only detected by consulting the source text in the Google Books collection. However, due to copyright restriction such double-checking is not always possible.

The above discussion shows that corpora such as the DTA, whose construction is quite labor-intensive, due to the amount effort required in double-keying, spelling normalization, linguistic annotation, and manual metadata creation have distinct advantages in data accuracy and reliability over the Google book corpus collections. However, this does not mean that the Google Books collections are irrelevant for diachronic linguistic studies. They are very useful as a secondary source of information that help to double-check the validity of results obtained from corpora such as the DTA.

5 Further Applications

While this study documents the ways in which certain words have waxed and waned in their use and frequency, with consequences for their meaning, there are wider implications that go beyond those important lexical details. In particular, the value of the corpora consulted and of the search tools they provide has clearly been demonstrated by the results that they allow for. At the same time, these results show that there are limitations on lexeme-based searches, in that our understanding of the developments that the *Alban(es)-* lexical items underwent crucially emerged from an examination of the context for each item, provided by the corpora and tools, disambiguating Italian *Albaner* from Balkan *Albaner*. These developments in turn provided some insight into mechanisms for semantic change viewed "up close" in a relatively short time span. Finally, it is a well-known problem in dealing with names of peoples and of groups that one and the same group can have multiple names in different, even related, traditions (e.g. *Deutscher*, *German*, *allemand*, etc.); this problem is acute in the case of group names from the distant past. The example of *Alban(es)-* shows how it is possible to untangle multiple names for the same referent through careful corpus searches and accompanying manual work. The ability to do so enhances, for instance, the prospects of undertakings like the Herodotos Project (`https://u.osu.edu/herodotos/`), aimed at developing a comprehensive listing of group names mentioned in Classical sources and modeling the networks of those groups.

In support of the Herodotos Project, Erdmann et al. (2016) employ a Named Entity Recognition system that identifies textual references to group names and the personal and place names with which they co-occur, but they have yet to disambiguate the nature and context of each reference. Like the *Alban(es)-* lexical items, there are many references in Ancient Latin or Greek corpora that can map to multiple concepts, meaning that one name could refer to any one of several groups, given the context. Conversely, there are many concepts that map to multiple references, and furthermore, the very identity of these concepts and the nature of these mappings can change over time, just as *Albaner* evolved from referring to an area near Rome to referring to the Albanian people. The same data-driven approach combined with manual analysis employed in this paper can elucidate such evolving reference-concept mappings, enabling projects like the Herodotos Project to better understand the relationships between the named entities it extracts from historical texts.

One example of a case study of interest to the Herodotos Project is provided by the term *Thebes* and references *Thebans* in Greek. *Theban*, of course, refers to people from Thebes, a geographically based designation. However, the geography is ambiguous in that there is a Thebes in ancient sources in Egypt

and one in Greece (specifically in the region called Boeotia), as well as a few others too, so that Greek Θηβαῖοι (Thēbaîoi) could in principle refer to people from either city and in fact any use of the stem Θηβα- (Thēba-) would be potentially ambiguous. Typically contextual information can disambiguate various instances of the stem Θηβα-. For instance, in Iliad 14.113-4, Diomedes says:

(4) πατρὸς δ᾽ ἐξ ἀγαθοῦ καὶ ἐγὼ γένος εὔχομαι εἶναι /
patròs d' ex agathoû kaì egō génos eúkhomai eînai /

Τυδέος, ὃν Θήβῃσι χυτὴ κατὰ γαῖα καλύπτει
Tydéos, hòn Thēbēsi khytē kata gaîa kalyptei

'I too can declare my stock to be from a noble father, Tydeus, whom the heaped earth in Thebes now covers'

In this case, the reference to the Greek hero Tydeus, within the immediate context here, serves to locate the referent of Θήβῃσι 'in Thebes/DAT.PL', as the Greek Thebes, not the Egyptian one. Similarly, reference in Homeric epic to "Thebes with one hundred gates" indicates a reference to the Egyptian city, not the Greek one.

There are many more cases like the Thebes case, since many colonies were named after the home city of the colonizers; for instance, there is a Κύμη (Kymē) in Euboea, an island just east of central Greece, but there is also one in southern Italy which was founded by settlers from the Greek city. Contextual cues such as those leveraged by the DiaCollo tool can similarly differentiate which place is referenced in such cases.

One final example comes from the opening of Caesar's *De Bello Gallico* in which he asserts that one of the groups inhabiting Gaul is known as *Celti* in their language, and *Galli* in Latin (Caesar, *BG* 1.1). In other words, Caesar not only maps two references to one concept, but also clarifies who is more likely to use which reference. In this case, gathering collocations for both references in a range of texts and analyzing the output would shed critical light on perspectives and biases before and after Caesar's campaign into Gaul. The sentiment of such collocations would demonstrate how *Galli*-reference users depicted the Gauls as compared to *Celti*-reference users. A diachronic analysis would address questions such as whether these biases converged or diverged after Caesar's campaign and how long convergence/divergence took in addition to other such questions. Perhaps the two references are actually better modeled as referring to two distinct but related concepts: the romanticized idea of *Celti* versus the barbarian caricature of the *Galli*. This stance can be supported or rejected with such an investigation, and developing a language-agnostic version of DiaCollo would greatly contribute to the feasibility and success of such investigations.

6 Conclusion

The methodology demonstrated here, combining automatic collocation analysis and manual inspection of the results can both identify and shed light on complex relationships between lexical items and the concepts they refer to, as well as how those reference mappings evolve over time. While DiaCollo enables this process to be very effective in German, a comparable language-agnostic tool will need to be developed in order to address many cases of interest in other languages without requiring potentially problematic translation. Regardless of the future of such technology, it is clear that homonymous named entities and entities with multiple names present a major challenge for attempts to automatically infer information about them. We present here a framework for addressing such challenges in a manner sophisticated enough to inform and impact qualitative research in the humanities.

Acknowledgements

The research of the first author was supported by the CLARIN-D grant provided by the German Ministry for Education and Research (BMBF). All authors express their gratitude to three anonymous reviewers for their detailed comments on an earlier version of this paper, which was presented at the CLARIN Annual Conference in Aix-en-Provence. Special thanks go to Bryan Jurish of the Berlin-Brandenburg Academy of Sciences for his extensive help with the use of the DiaCollo tool.

References

[Bloomfield1914] Leonard Bloomfield. 1914. *Introduction to the Study of Language*. Henry-Holt, New York.

[Bloomfield1933] Leonard Bloomfield. 1933. *Language*. Henry-Holt, New York.

[Davis2012] Mark Davies. 2012. Expanding Horizons in Historical Linguistics with the 400 Million Word Corpus of Historical American English. *Corpora*, 7:121–157.

[Erdmann et al.2016] Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner and Marie-Catherine de Marneffe. 2016. Challenges and Solutions for Latin Named Entity Recognition. *Proceedings of the Language Technologies for the Digital Humanities Workshop* in conjunction with the *26th International Conference on Computational Linguistics* (COLING-2016), December 2016.

[Geyken2007] Alexander Geyken 2007. The DWDS Corpus: A Reference Corpus for the German Language of the 20th Century. C. Fellbaum ed. *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. Bloomsbury Academic, London. p. 23-41.

[Geyken et al.2011] Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas und Frank Wiegand. 2011. Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. S. Schomburg et al. eds. *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*. pp. 157-161.

[Haaf et al.2013] Susanne Haaf, Frank Wiegand, and Alexander Geyken 2013. Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text. *Journal of the Text Encoding Initiative* (jTEI) 4.

[Hinrichs and Krauwer 2014] Erhard Hinrichs and Steven Krauwer 2014. The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC-2014), May 2014, pp. 1525–31.

[Hinrichs and Trippel in press] Erhard Hinrichs and Thorsten Trippel in press. CLARIN-D: eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften. *Bibliothek Forschung und Praxis*; Vol. 41.1 (April 2017).

[Hock and Joseph1996] Hans Henrich Hock and Brian Joseph. 1996. *Language History, Language Change, and Language Relationship. An Introduction to Historical and Comparative Linguistics*. Mouton de Gruyter (2nd edn., 2009), Berlin.

[Jurish et al.2014] Bryan Jurish, Christian Thomas, and Frank Wiegand. 2014. Querying the Deutsches Textarchiv. In: U. Kruschwitz, F. Hopfgartner, and C. Gurrin eds.: *Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities* (co-located with iConference 2014, Berlin, 4. März, 2014), p. 25–30.

[Jurish2015] Bryan Jurish. 2015. DiaCollo: On the Trail of Diachronic Collocations. K. De Smedt ed. *Proceedings of the CLARIN Annual Conference 2015*. Wrocław, Poland, 15th-17th October, pp. 28-31.

[Libelt1828] Karol Libelt. 1828. *Wykłady Humboldta na uniwersytecie Berlińskim: notaty prelekcyj tych po uczniu Jego Karolu Libelcie* [= Nachschrift der 'Kosmos-Vorträge' Alexander von Humboldts in der Berliner Universität, 3.11.1827–26.4.1828].

[Lieberman et al.2007] Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin Nowak Quantifying the Evolutionary Dynamics of Language. *Nature* 449 (2007).

[Michel et al.2012] Jean-Baptiste Michel, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Lieberman Aiden 2012. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, DOI: 10.1126/science.1199644.

[Mommsen1854] Theodor Mommsen. 1854. *Römische Geschichte. Bd. 1: Bis zur Schlacht von Pydna*. Leipzig, Germany.

[Rychlý2008] Pavel Rychlý. 2008. A Lexicographer-friendly Association Score. *Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Language Processing RASLAN 2008*, pp. 6–9.

[Schiller et al.1995] Anne Schiller, Simone Teufel, and Christine Thielen. 1995. *Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technical Report. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, and Seminar für Sprachwissenschaft, Universität Tübingen.

[Zhang2015] Sarah Zhang. 2015. The Pitfalls of using Google Ngram to study Language. *Science* 10.12.15.