What's in a Name? The case of *Albanisch-Albanesisch* and Broader Implications

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

Erhard Hinrichs
Department of Linguistics
Eberhard-Karls Universität
Tübingen, Germany
erhard.hinrichs@uni-tuebingen.de

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

Abstract

This paper offers a use case of the CLARIN research infrastructure 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 tools that are available for the DTA and the DWDS corpora as part of the CLARIN-D infrastructure 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 southeastern 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 for the people and the language: *Albanisch-* and *Albanesisch-*, and even in English, in one author, linguist Leonard Bloomfield (Bloomfield 1914, 1933), the variant *Albanese* for the language name is encountered.

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 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 German books (henceforth: GBCG), the Deutsche Text Archiv Archive (henceforth: DTA; Geyken et al. 2011; www.dta.de), and the corpus of the Digitales Wörterbuch der deutschen Sprache (www.dwds.de), 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; Davies 2012) and the Google books 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. Section 4 summarizes the results and discusses some broader implications of the present use case.

2 Comparative Study of Historical Corpora for English and German

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 German

Figure 1: Search results for *Albanisch*/*Albanesisch*

Fig. 1 shows the results for all word forms of *Albanisch-* and *Albanesisch-* for the GBCG. *Albanesisch-* outranks *Albanisch-* in relative frequency between 1830 and 1915 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. However, the search results for the nouns *Albaner-* and *Albanese-* in Fig. 2 differ from the results in Fig. 1 in that the former outranks the latter for entire period covered by the GBCG.

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

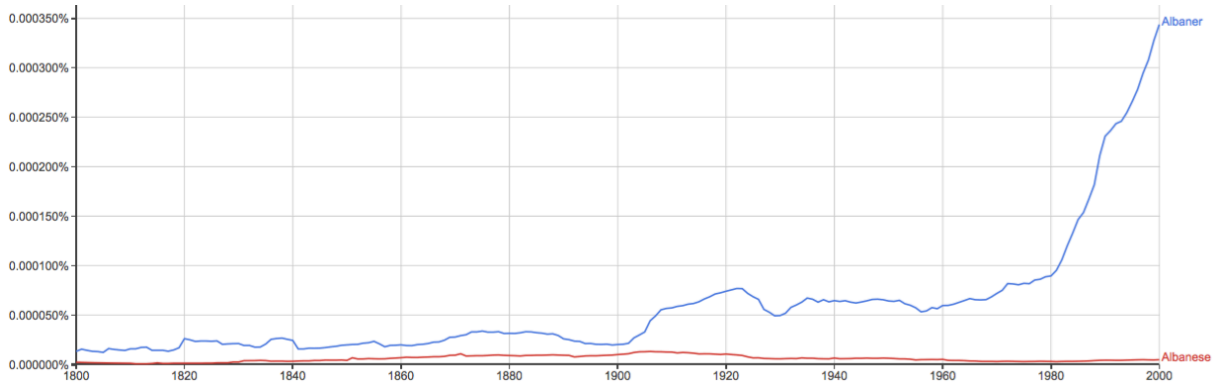

Figure 2: Search results for *Albaner*/*Albanese*

Are we to conclude from Fig. 2 that *Albaner* was the preferred term of reference for persons from Albania, with *Albanese* a secondary variant? The mere frequency counts in Fig. 2 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 Book n-gram 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.

Lemma	Frequency Count	Earliest Data Point	Latest Data Point
Albanisch	97	1650	1913
Albanesisch	19	1789	1913
Albaner	61	1663	1913
Albanese	36	1789	1913

Table 1: DTA query results

Inspection of the linguistic contexts for all data points DTA data 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* refer to persons from Albania. By contrast, all instances of the lemma [a|A]lbaner- in the DTA refer to people or locations north 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 its uses referring to entities related to the territory north of Rome and other instances referring to entities

related to Albania. The use of the terms *Albanesisch-* and *Albanese-* may therefore at least partly be motivated by the well-attested and well-motivated pragmatic strategy of trying to avoid ambiguity.

Figure 3: Bigram Comparison of *albanische Sprache* und *albanesische Sprache* in the GBCG

Further evidence for this strategy being put to use can be gleaned from the comparison of *albanesische Sprache* and *albanische Sprache* in the GBCG shown in Fig. 3. Notice that the former has higher frequency up until usage of the unigram *albanesisch* started to decline; at that point, there was no longer a need for avoiding ambiguity and accordingly there is a decline.

3 Tracking Semantic Change in the DTA and DWDS Collections

The CLARIN-D tools and resources available at the BBAW also help to trace the historical change from referring to people from Albania as *Albanesen* to the contemporary term *Albaner-*. The web application DiaCollo collects sets of collocates for a given word for regular time intervals. Changes in collocation behavior of the target word are one diagnostic for changes in word meaning over time. Table 2 shows that the set of noun collocates identified by DiaCollo for the lemma *Albaner-* and the time period for which they are characteristic. The DiaCollo visualization allows easy inspection of the corpus evidence in keyword-in-context format for each occurrence of a collocate.

Collocate Noun	From	To
Römer 'Roman'	1670	1888
Gebirge 'mountain range'	1716	1977
Stein 'stone'	1716	1977
Berg'mountain'	1910	1988
Serbe 'Serbian'	1982	1990
Provinz 'province'	1982	1990

Table 2: DiaCollo Results for the Collocates of the lemma *Albaner* during 1600-1990

Examination of the linguistic contexts of the collocates reveals that with the exception of *Serbe* and *Provinz*, where *Albaner-* refers to persons from Albania, all other collocate nouns refer to the Italian region north of Rome or to the Roman adversaries of the Alban people. The change in meaning of the term *Albaner* is, thus, a fairly recent phenomenon, coinciding roughly with the Balkan wars in the 1990s. 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 & Joseph 1996: 224), may have been at work here.

4 Conclusions and Wider Implications

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 in secondary literature on those sources.

Reference

Bloomfield, L. 1914. *Introduction to the Study of Language*. New York: Henry Holt.

Bloomfield, L. 1933. *Language*. New York: Henry Holt.

Davies, M. 2012. Expanding Horizons in Historical Linguistics with the 400 million word Corpus of Historical American English. *Corpora* 7, pp. 121-57.

Geyken, A., S. Haaf, B. Jurish, M. Schulz, J. Steinmann, C. Thomas and F. Wiegand. 2011. Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In: S. Schomburg et al. *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*, pp. 157-161.

Hock, H. H. and B. Joseph. 1996. *Language History, Language Change, and Language Relationship*. *An Introduction to Historical and Comparative Linguistics.* Berlin: Mouton de Gruyter (2nd edn., 2009).

Jurish, B. 2015. DiaCollo: On the trail of diachronic collocations. In K. De Smedt (editor), *Proceedings of the CLARIN Annual Conference 2015*. Wrocław, Poland, 15th-17th October, pp. 28-31.