Towards a Spatial Analysis of Toponym Endings

Tobias Dahinden

Abstract The target of this article is to define and use a statistical measure to determine endings of place names. The definition of 'ending' is based on the occurrence of a certain end-string in a gazetteer. Based on this definition a part of the GeoNames-gazetteer is analysed in respect to detect and rank possible endings. The spatial distributions of the most outstanding endings are presented.

Keywords Toponymy · Analysis · Detection · Gazetteer · Vernacular

1 Introduction

Toponyms consist of a single name (simplex), composition and move together (compound), and derivation (derivata). Similarities between toponyms are obvious: some places use the same name (e.g. *Cambridge*) and, at least in Germanic languages, some names of places have the same ending.

Concerning endings, it is possible to differ between suffixes (e.g. *-nia* in *California* and *Pennsylvania*) and primary words (e.g. *-land* in *Greenland* or *Maryland*).

The aim of this article is to detect salient endings of Toponyms in a statistical way, and to present the source area for some of them. The article has been organized in the following way: in the Sect. 2 some knowledge about endings is presented. A statistical definition of 'ending' is given in Sect. 3. The Sect. 4 gives a brief overview on the gazetteer that is used in this research. Some results are presented in Sect. 5, i.e. endings that are salient according to the definition in

T. Dahinden (✉)
Institut für Kartographie und Geoinformatik, Leibniz Universität Hannover, Hanover, Germany
e-mail: tobias.dahinden@ikg.uni-hannover.de

M. Buchroithner et al. (eds.), *Cartography from Pole to Pole*,
Lecture Notes in Geoinformation and Cartography, DOI: 10.1007/978-3-642-32618-9_26,

Sect. 3 and the source area of some endings are shown. The Sect. 6 gives an outlook on future work.

2 Expectations about Endings

A considerable amount of literature has been published about names and endings of names, mainly in linguistics. In general, an author investigates a small area and analyses the names based on the meaning of parts of the names. As an example, Andrießen (1991) traced for Hessen (Federal state in Germany) *"-berg, -hagen, -hausen, -inghausen, -rode* [and] *-sen (< -hausen)*" as common endings for place names. Waser (2002) reported on place names from the German-speaking part of Switzerland. He discussed the diminutives *-i, -etli, -ti, -li, -ili, -eli, -dli* in place names. According to a publication of Bickel (1998) the common endings of toponyms for north-west Switzerland are *-walen, -ingen, -berg, -wil, -inghofen, -heim,* and *-dorf*.

From the referred research it can be noted that

- the size of the endings differs from 1 to 9 characters (with a mean value of 4),
- endings are typical for certain regions,
- some endings are a superset of another ending, such as *-sen* for *-hausen*.

A long but unproved list with endings consisting on suffixes and primary words is published on Wikipedia (2012).

Names may have a misleading meaning in the current spoken language. According to Bossong (2002) toponyms are usually adapted to the current language although they are often much older. He mentioned a place called *Bischofsheim*[1] in the Vosges Mountains were at no time a "Bischof" was present. He mentioned that there are similarities to the toponym *Bizcoy* (a place near Alicante in Spain). A related toponym is *Biscofesheim*, which is the old name for *Tauberbischofsheim* (Wikipedia 2009). So, the hyphenation of *Bischofsheim* should be rather *Bisc-hof-(s)-heim* than *Bi-schof-(s)-heim*. It would be too easy to compare the names with a dictionary of the language of this area.

The purpose of this research is to find patterns in endings of toponyms in a gazetteer with entries that are widely distributed. Because analyses on endings are usually for a rather small area it is not satisfying just to apply the known endings of some areas to a larger region. It is also not sufficient to use a dictionary to find primary words because names are sometimes much older than the spoken language.

Thus, we suggest using a statistical method together with a gazetteer instead of using an existing list of common endings. Unfortunately, there is no statistical definition for endings according to our knowledge.

[1] Translated: the Bishop's home.

3 A Statistical Definition for Endings

As a simplification for this article, the term ending is used for a suffix as well as for a primary word or a compound. To determine common endings in a gazetteer a statistical definition is unavoidable.

An ending could be defined as something like an end-string *s* of a word, which occurs more often than other end-strings. It has to be taken into account that longer strings are less often than short strings because a long string is always a subset of a short string. Hence, endings cannot be defined by the number of occurrence only; otherwise an end-string like *-orf* would be a more prominent ending than the primary word *-dorf*.

The definition might be based on the idea of checking weather adding an additional character to a certain substring could be random or systematic. Thus, the occurrence of a character *c* in front of a string *s* is analysed. This is done using the ratio *R* of the number of words with the ending $c + s$ and the number of words with ending *s*. Shortly written: $\#(c + s)/\#s = R$.

The chance *C* of an end-string to be an ending is regarded as the sum of the occurrence-ratio *R* of the letters divided by the length of the substring.

Def. An ending is an end-string *s* where the chance *C* of being an ending is greater than *C* of all substrings, which are longer or shorter by one character.

As an example the string *-nplanggen* is discussed. It's the most often occurring end-string in a gazetteer called *Urner Namenbuch*[2] (Hug and Weibel 1988–1991). The end-string *-nplanggen* consists of nine characters and thus, can be subdivided in nine elements (c.f. Table 1). It occurs in 130 names while *-planggen* occurs in 534 names. This leads to $R = 0.24$. This means that 24 % of the words with the substring *-planggen* also have the substring -nplanggen. The substring *-langgen* occurs as often as *-planngen*, thus *R* is 1.

The chance *C* for *-nplanngen* to be an ending is the sum of all ratios *R* ($5.97 = 0.24 + 1 + \ldots + 0.26$) divided by the number of characters (9). An analysis of Table 1 shows that a maximum for *C* occurs for the string *-planggen* and for *-en*. Actually, the term *planggen* is a primary word (derived from latin *planca*) in the language represented in the gazetteer.

4 Gazetteer and Research Restrictions

This study makes intensive use of the GeoNames gazetteer (Wick 2012). The gazetteer is based on more than 60 different data sources such as Wikipedia's Wikipedia-World, U.S. Geological Survey's Geographic Names Information System, and Swisstopo's Swissnames. It contains more than 8.2 million

[2] Digital access on: www.ortsnamen.ch.

Table 1 Division of *-nplanngen* in possible endings: *R* gives the ratio of counts of a word with its predecessor. *C* is the sum of *R* of the letter and all predecessors divided by the length of the string, large values indicate that it is very improbable that this term occurred by accident

Ending	Counts	*R*	Sum(*R*)	Length	*C*
-nplanggen	130	0.24	5.97	9	0.66
-planggen	534	1	5.73	8	0.72
-langgen	534	1	4.73	7	0.68
-anggen	534	0.95	3.73	6	0.62
-nggen	562	0.73	2.78	5	0.56
-ggen	762	0.71	2.04	4	0.51
-gen	1067	0.16	1.32	3	0.44
-en	6596	0.91	1.17	2	0.58
-n	7288	0.26	0.26	1	0.26

geographical names; it is published using the Creative-Commons Attribution 3.0 License. It stores for each place a name, a latitude and longitude in WGS84, an id, a name in ASCII, a list with alternative names, information about the feature type, population, the code about the administrative division, and the elevation.

Some names in the gazetteer consist of more than one word (such as "Moreira de Cónegos" or "Museum of Modern Art"). In such cases it is not possible to know a priori which part should be analysed. Thus, for the study the names were restricted to consist of single words.

The investigation area is restricted to Germany, Austria, and Switzerland. The majority of names are from an area where German is spoken. This means that the research is restricted to agglutinative languages; these are languages where compounds are set together. Thus, the words can be very long.

These restrictions shortened the list with places to 178032 names. The mean density of the entries in the database is around one name per 2 km^2 for this area.

The analysis was performed on the names written in ASCII, no phonetics was used. Thus, Umlauts were not taken into account,[3] two very common digraphs (*sc,*[4] *ch*) were handled as one single character; other digraphs (*ck, ie, ph, tz,*…) are handled as two characters. The length of the endings is restricted to 6 letters to reduce the runtime of the analysis.

The result is restricted to substrings that occur at least 100 times in the gazetteer. Still, there are 1013 different end-strings to be analyzed.

[3] Unfortunately, Umlauts such as *ü* were stored inconsistent, partly as *u* and partly *ue*.

[4] This is related to *sch.*

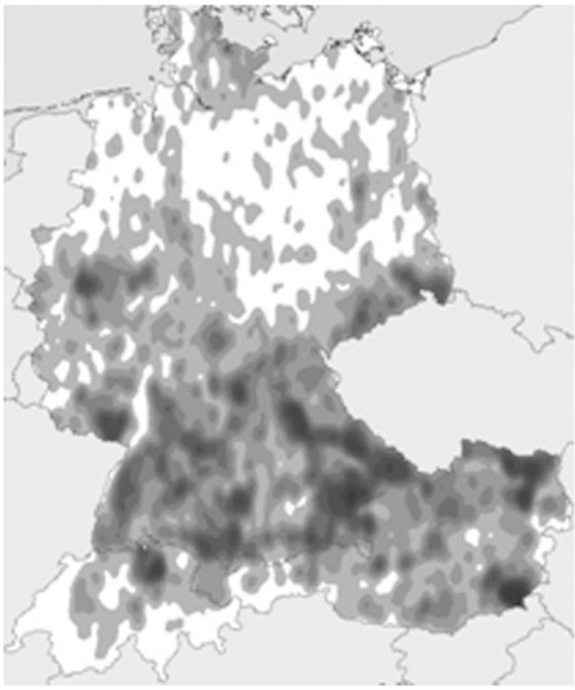

Fig. 1 Distribution of places with a name ending on *-berg*

5 Analysis

Using the definition of Sect. 3, the end-strings of the place names are classified and ranked. The 12 strings with the highest chance C to be an ending are: *-berg, -spitze, -enberg, -hausen, -dorf, -weiler, -endorf, -erberg, -kreuz, -graben, -enkamp, -heim*.[5] Comparing this list with literature about endings (c.f. Sect. 2), it becomes obvious that the evaluation method privileges primary words and compounds of *-en* or *-er* with a primary word, e.g. *-kamp*.

Figures 1, 2, 3, 4 show the probability density function of places with the ending *-berg*, *-dorf*, *-weiler*, and *-enkamp*. The images are based on kernel density estimation. All of them have been calculated using an Epanechnikow kernel (ESRI Support Centre 2010) with a bandwidth of 0.2°. The bandwidth has been established by a visual observation.

As one can see from the Figures the distribution varies for different endings. While places ending on *-berg* exist nearly everywhere in the investigation area, places ending on *-dorf* are common in the eastern part, *-weiler* is common in the shout-west and *-enkamp* seems to be typical for the north-west of Germany.

[5] A longer list is presented in the appendix of this article.

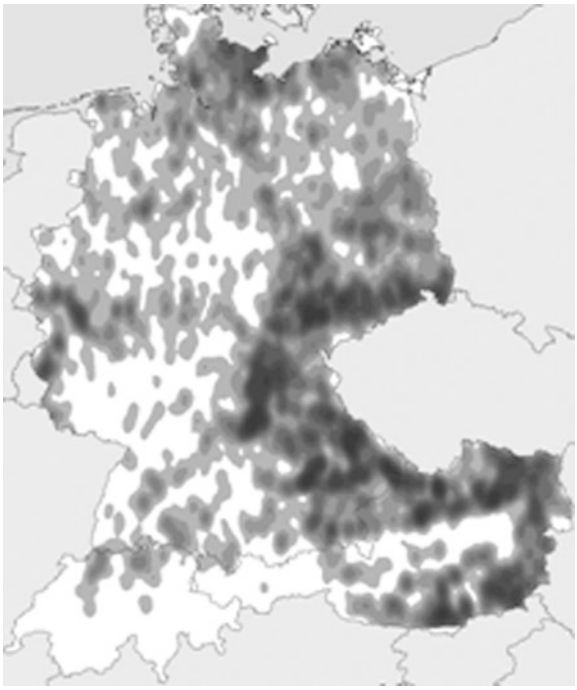

Fig. 2 Place names ending on *-dorf* occur more frequently in the eastern part of the investigation area

Fig. 3 The ending *-weiler* seems to be related to the south-west

Fig. 4 The end-string -*enkamp*, that is a compound of the suffix -*en* and the primary word -*kamp*, is common in north-west Germany

Figures 5 and 6 show the density estimation of 42 endings. The estimation is based on a Nadaraja-Watson type kernel smoother (Baddeley and Turner 2005, Diggle and Arnold 2003) with a bandwidth of 1°. The endings were selected according to *C*. Compounds like -*enkamp* are not taken into account for the presentation. In Fig. 5 the colour scale is the same for all images while it changes in Fig. 6.

Figure 5 is useful to uncover endings that dominate a region, noticable are: -*au,* -*bach,* -*berg,* -*dorf,* -*feld,* -*hausen,* -*hof,* -*ingen,* and -*ow*. Unsurprisingly, these are the same endings that occur most often in the database.

Figure 6 is useful to see, where the endings are probably located. Comparing the single images correlation between endings could be guessed, e.g. there might be a correlation between -*horst*, -*moor*, and -*stedt*. What is surprising is that there could be a kind of negative correlation of -*dorf,* -*hausen,* and (maybe) -*heim*.

Figure 7 provides the distribution of -*ingen.* The image was calculated using adaptive kernel density estimation (Silverman 1986, Baddeley and Turner 2005) with 20 repetitions. From the Figure it becomes obvious that this ending is present mainly in the South but also in northern regions.

Comparing the endings to dialect maps lead to some correlation between the endings and the dialects, e.g. -*ingen* is often used where people speak Alemannic German (south west) but it is also appearing in northern parts of Germany. On the

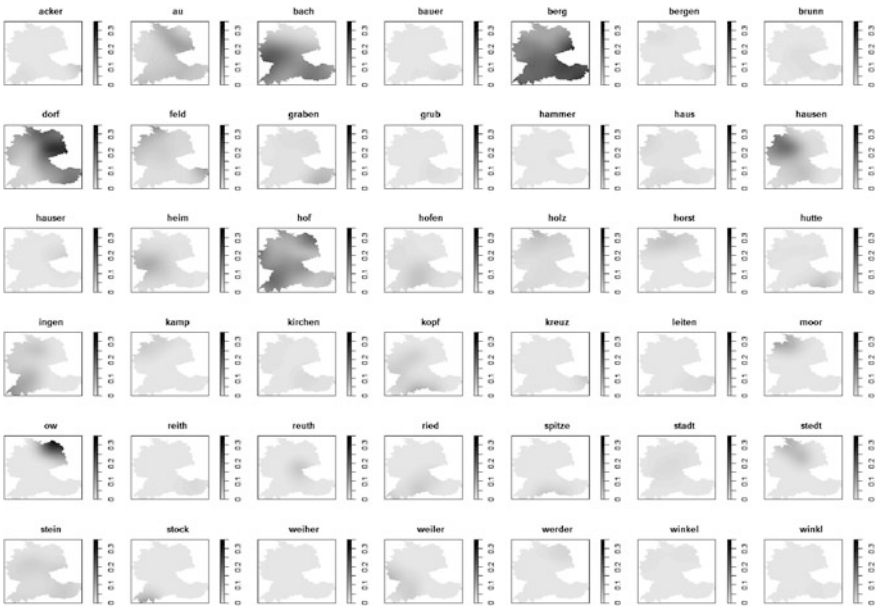

Fig. 5 Density of 42 different endings (the darker the area, the more probable an ending occurs). All images are displayed using the same scale

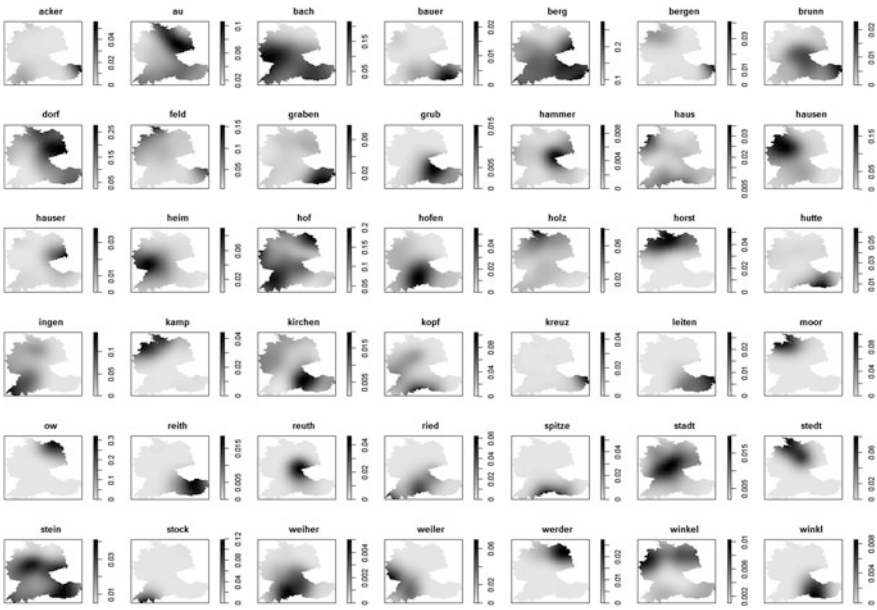

Fig. 6 Density of 42 different endings (the darker the area, the more probable an ending occurs). Note: The images use different scales

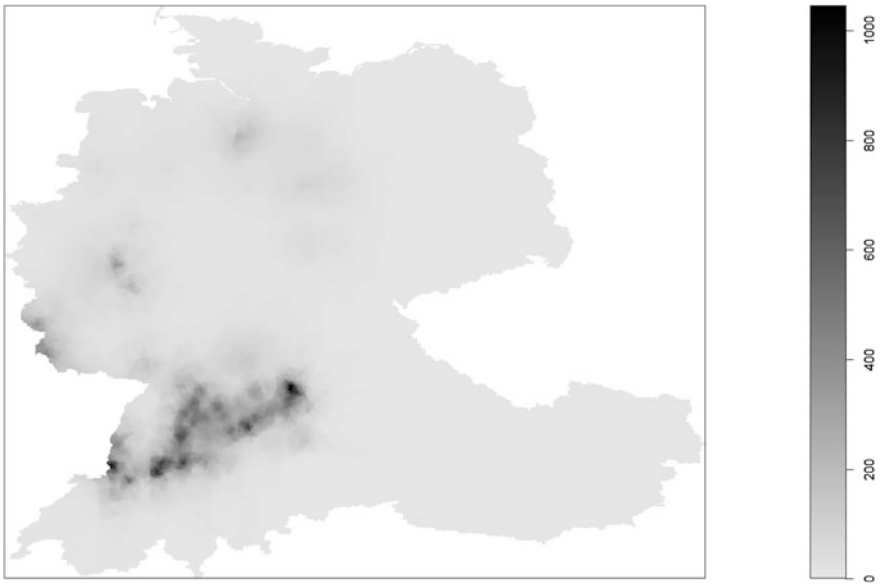

Fig. 7 Probability to find places ending on *-ingen* (the darker the area, the more probable the ending occurs). The image has been calculated using adaptive kernel density estimation

other hand, there might be a correlation between *-horst, -moor, -stedt* and Lower German.

6 Conclusion and Further Work

In the article the term 'ending' has been defined using a measure based on the absolute number of end-strings in a database and the length of the string. The measure is rather simple; e.g. it does not take into account that letters have a certain frequency.

The definition of 'ending' does not distinguish between primary words, suffixes, simplex, and compound. Future studies are recommended to deal with the splitting of endings. This would involve the use of a dictionary with primary words. The dictionary itself could be established using the definition of this paper. On the same time, endings should not be restricted in length, unlike it has been done here. Further, the analysis should be enlarged to languages where words tend to be isolated. This implies that the analysis must not be restricted to endings but it has to deal with any kind of patterns in words.

For the analysis of the place names the investigation area was restricted to German-speaking countries and only ASCII names of the GeoNames-database were used. A more profound study could deal with phonetic algorithms such as

Soundex (c.f. Knuth 1973) or *Kölner Phonetik* (Postel 1969). Using such a method would allow bundling similar words. The German *-heim* and the English *-ham* or the German *-burg* and the French *-bourg* would be handled correctly as the same word. Similar the German *-reuth, -rode, -reit, -rütti, -ried, -roda,…* would be seen as one single expression. But undesired friends may also arise: the Kölner Phonetik for *Othenbruch, Osnabrück, Oschenberg* is *O86174*, meaning the words are the same according to the algorithm, but most probably they are not. A solution would be the use of the International Phonetic Alphabet as a base for an analysis.

The analysis was performed on a small part of the GeoNames-database and a ranking of possible endings of place names has been created based on the definition in Sect. 3. The spatial distribution of the most-outstanding endings has been shown. A visual analysis of the distribution suggested weak correlation between endings, possibly also negative correlation. The current study was unable to make a profound analysis on spatial point patterns. It is suggested that a detailed statistical analysis on the distribution is applied. A major issue would be the un-equal distribution of the places in space and to run a multivariate point pattern analysis related to the end-strings.

Further studies on the current topic should deal with larger investigation regions and/or more detailed gazetteers. Research questions that could be asked include the existence of ending structures in other languages, the correlation of patterns in place names and the spatial distribution of these names, the analysis of compounds, the search for pattern in parts of the names, and an analysis of names consisting of more than one word.

A.1 7 Appendix

The 150 salient endings evaluated for this article (beside those starting with *-en*, *-er*, or *–es*):

-berg, -spitze, -hausen, -dorf, -weiler, -kreuz, -graben, -heim, -weiher, -kamp, -hauser, -bach, -grub, -elberg, -stein, -reuth, -schlag, -winkl, -acker, -holz, -chberg, -scheid, -hammer, -feld, -neuhof, -werder, -brunn, -nsdorf, -hofen, -stedt, -kopf, -reith, -stadt, -ndberg, -winkel, -tzberg, -haus, -horst, -rnberg, -ingen, -hof, -bauer, -stock, -onberg, -leiten, -ried, -au, -msdorf, -nsberg, -bergen, -gsdorf, -ow, -hutte, -eldorf, -moor, -ssberg, -strass, -garten, -isberg, -undorf, -storf, -indorf, -tsberg, -trup, -itz, -dsberg, -amberg, -wiesen, -hlberg, -spitz, -krug, -heide, -bruck, -moos, -holzen, -lsheim, -brink, -llberg, -steig, -itze, -burg, -schach, -grund, -berge, -muhle, -aubach, -koog, -kogel, -elbach, -schutz, -usen, -wald, -chholz, -bronn, -sgrun, -en, -kanal, -hagen, -hardt, -grat, -statt, -chfeld, -reute, -heid, -tzbach, -hofe, -reit, -iler, -hlag, -thal, -er, -buhel, -muhlen, -lehen, -wiese, -felden, -halden, -inbach, -bichl, -ppach, -bruch, -nfels, -rnbach, -riegel, -eck, -roth, -mmer, -sleben, -imbach, -elhof, -sattel, -ikon, -etten, -aben, -ambach, -buttel, -hafen, -hutten, -ewitz, -oven, -ing, -harte, -ssbach, -hub, -asser, -hlbach, -roda, -acher, -furth, -stall

References

Andrießen K (1991) Siedlungsnamen in Hessen. Verbreitung und Entfaltung bis 1200. Deutsche Dialektgeographie. N.G. Elwert, Marburg

Baddeley A, Turner R (2005) Spatstat: An R package for analyzing spatial point patterns. J Stat Softw 12(6):1–42

Bickel H (1998) Ortsnamen als Quellen für die Siedlungsgeschichte am Beispiel der Nordwestschweiz. XIXth international congress of onomastic sciences, Aberdeen, 4–11 Aug 1996

Bossong G (2002) Der Name Al-Andalus: neue Überlegungen zu einem alten Problem. In: Restle D, Zaefferer D (eds) Sounds and systems. Studies in Structure and Change. A Festschrift for Theo Vennemann. Mouton de Gruyter, Berlin, pp 149–164

Diggle PJ, Arnold H (2003) Statistical analysis of spatial point patterns. Arnold, London

ESRI Support Centre (2010) ArcGIS desktop help: kernel density. Environmental Systems Research Institute, Inc. http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?id=6190&pid=6188&topicname=Kernel_Density. Accessed 13 April 2010

Hug A, Weibel V (1988-1991) Urner Namenbuch: Die Orts- und Flurnamen des Kantons Uri. 4 vols. Altdorf

Knuth DE (1973) The art of computer programming: volume 3. Sorting and searching. Addison-Wesley, Reading

Postel HJ (1969) Die Kölner Phonetik - Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. IBM-Nachrichten 19:925–931

Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall (CRC Press), London

Waser E (2002) Das Diminutiv in Orts- und Flurnamen. Congreso Internacional de Ciencias Onomásticas, Santiago de Compostela

Wick M (2012) GeoNames. www.geonames.org. Accessed 8 Nov 2012

Wikipedia (2009) Tauberbischofsheim. Wikipedia Die freie Enzyklopäsie, Mediawiki Foundation. http://de.wikipedia.org/wiki/Tauberbischofsheim. Accessed 4 Oct 2009

Wikipedia (2012) German placename etymology. Wikipedia the free encyclopedia, Mediawiki Foundation. http://en.wikipedia.org/wiki/German_placename_etymology. Accessed 14 Nov 2012