In Search of Naming Patterns: A Survey of Finnish Lake Names

Antti Leino
University of Helsinki
antti.leino@cs.helsinki.fi

Abstract

The existence of patterns as one of the factors in the toponomastic process has been known for more than a quarter of a century. However, while some onomasticians have suggested that such patterns can play an important role even when the names in question can be adequately explained by other means, such hypotheses have been rather difficult to prove. The present study is an attempt to address the issue: the goals were, first, to find regularities in the naming of Finnish lakes; second, to assess whether such regularities imply the presence of naming patterns; and third, to see if a quantitative study could give new insights about the properties of such patterns.

This was done by applying methods developed in the computer science field of data mining to an electronic corpus consisting of all Finnish lake names found on the 1:20 000 Basic Map. These revealed several groups of names that appear next to each other significantly more often than could be expected, even after accounting for regional variation in the distributions of the names.

Some of the groups can be explained by referring to e.g. cultural history, but in a large number of groups the names have a semantic relationship which suggests that there is a large number of relatively widespread patterns in naming Finnish lakes. However, these patterns are very specific and it is difficult to see a systematically productive general pattern. Some of the phenomena involved can be described using Construction Grammar, but it is evident that the theoretical framework needs some adjustments.

Introduction

It is widely accepted that one of the contributing factors in the process of naming places is the use of existing patterns. Some onomasticians, e.g. Pamp (1991) and Kiviniemi (1977), have suggested that such patterns can play an important role even when the names in question can be adequately explained by other means. However, such hypotheses have been rather difficult to prove, and indeed it seems likely that the strongest form of such hypothesis is not provable: it is in general very hard to show conclusively that a given instance has resulted from such a process.

However, it is possible to show that a general tendency towards using analogy exists even with names that are clearly related to physical characteristics of the places. The present study started as an attempt to do just this, by searching for regularities in the naming of Finnish lakes and subsequently assessing whether such regularities imply the presence of naming patterns. This was done by applying methods developed in the computer science field of data mining to an electronic corpus of Finnish lake names.

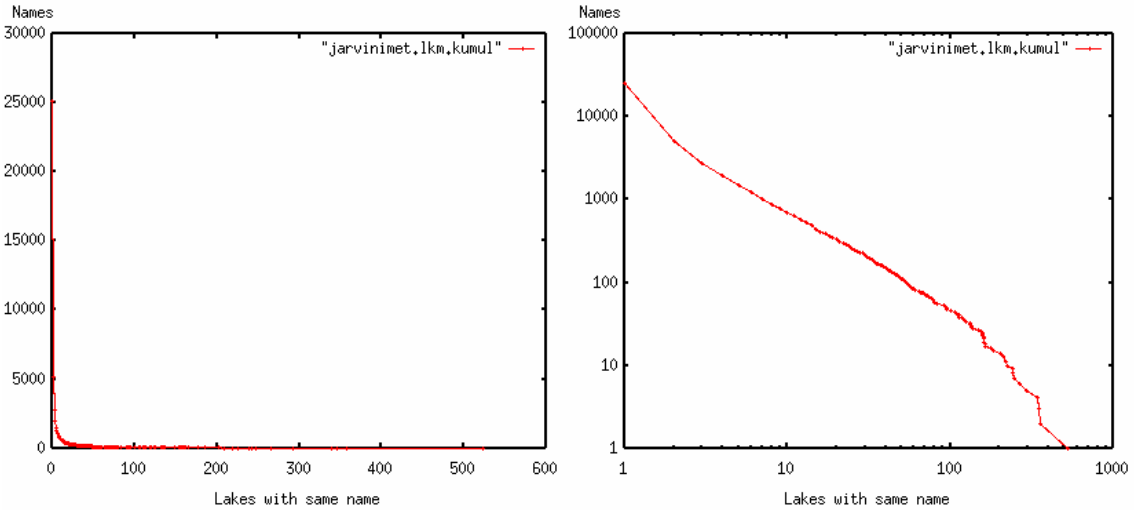
Data and Methods

The data used in this study comes from the Place Name Register of the National Land Survey of Finland (Leskinen 2002), used to produce the 1:20,000 Basic Maps. To narrow the scope, I chose all lake names that occur at least 20 times; the number of names in the entire register as well as the current selection is shown in table 1.

	Occurrences	Names	Named places
All toponyms	≥ 1	303 626	717 747
Lakes	≥ 1	25 178	58 267
This study	≥ 20	331	19 230

Table 1: Size of the place name corpus

The primary reason for selecting this subset was that the methods used — or, in fact, any meaningful quantitative study of interactions between two names — require that there are several occurrences of each name. Another reason was that the number of different names increases rapidly as the number of occurrences for each name decreases: as shown in figure 1, the numbers appear to follow the normal Zipf law. Setting the limit at 20 occurrences reduced the number of different names sufficiently to manage the corpus on a desktop computer.



Linear scale Logarithmic scale

Figure 1: Number of similarly-named lakes

The methods used to analyse this corpus are mostly based on the computer science field called data mining (e.g. Hand et al. 2001). The general goal in this field is to search for regularities in massive corpora of data whose structure is not necessarily well known in advance. As data mining methods have only recently been extended to spatial data, the methods used are also to a large extent based on previous work in statistics on spatial data analysis (e.g. Bailey—Gatrell 1995).

In terms of spatial statistics, the occurrences of each toponym can be viewed as a point pattern — that is, a set of points on a map. The small-scale interactions between different names then become what a statistician would call second-order effects[1] between two such point patterns; analysing such effects is a well-known problem, and there are established tools for the task. The primary one used in this study is a variation of the so-called *K* function (Ripley 1976). This can be considered as a distortion of area: in the original case, involving only one set of points, one starts with counting the number of points within a given radius *r* of any of the points. Now, *K(r)* shows how large an area around each point one would need to get this number of points if the points were actually randomly distributed.[2]

[1] In terms more familiar to onomasticians, the statistical concept of first-order effect corresponds to the overall distribution, whereas second-order effects are related to the influence a name has on the names of neighbouring places.

[2] More precisely, the expected number of points within radius *r* of another point is $E(n_r) = \lambda K(r)$, where λ is the overall intensity, or mean number of points per unit area.

If the spatial distribution of the points was completely random, *K(r)* would be the area of a circle with a radius of *r*, that is, πr^2. If, however, the points are attracted to each other, the *K* function is larger than this; if the points repel each other it is smaller. The *K* function is thus useful as a measure of the attraction between points. The variant used in this study is slightly more complex, in that it involves two different point patterns (that is, occurrences of two different lake names) and attempts to take into account that neither of these has a uniform overall distribution.

With this measure of spatial attraction it is possible to use variants of common data mining tools to find interesting groups of names. In this vein, this study has used a variant of a method called `Apriori` (Agrawal—Srikant 1994, Mannila et al. 1994), normally used for finding frequently occurring sets in large non-spatial data collections. Briefly, the method is based on the fact that if a set of items is frequent then all of its subsets must be at least equally frequent, so in searching for larger sets one needs only check combinations of already-found frequent sets. In the current case the task is to find groups of names that are attracted to each other, and it is obvious that in such groups all the sub-groups must also fulfil the same criterion. This is essentially similar to the premise of `Apriori`, and it is relatively easy to adapt the algorithm to the task in hand.

Experiment and Results

In general, there are no hard and fast rules on how to set the parameters in this sort of data mining experiment: the consensus is that one should try different values and see whether these give in interesting results. This is admittedly rather vague, and data mining methods can easily be misused to get results that have little basis in reality. On the other hand, in the current case an approach like this was justifiable, as prior studies (Leino et al. 2003) had already established that there are indeed statistically significant regularities in the data.

In the end I searched for groups of names where the inter-name *K* function at the radius of 1km was more than 20 times the theoretical value expected for spatially random data. A small radius like this means that the search concentrated on relatively small-scale interactions; this seemed necessary, as the goal was to study the effect of analogy in naming neighbouring lakes. Setting the cut-off at 20 times the theoretical value gave a relatively large, but still manageable number of co-occurring groups.

The mining experiment resulted in several groups of names that fit the criterion — that is, there was relatively strong attraction between each of the members of such group in the 1km range. The number of groups is shown in table 2; the numbers in the final column do not match the total, as there were overlapping groups of three and four elements. Practically none of the three- and four-member groups were interesting as such, but they often included as some sort of nucleus a pair that was.

Size of group	Number of groups	Number of distinct pairs
4	2	12
3	104	255
2	638	638
Total	744	903

Table 2: Attraction groups found

By "interesting" I mean a group where the reason for the spatial attraction can be seen. In the majority of the groups this was not the case, which indicates that there are other factors in name-giving than patterns and analogy. This shouldn't be a surprise.

Some of the attraction groups seem to result from cultural phenomena or the overall characteristics of the terrain. It is natural to suggest that similar agricultural environment tends to suggest similar names. Thus, to take as an example one of the attraction groups, an environment that gives rise to *Niittylampi* 'Meadow Pond' could very well also find motivation for *Vasikkalampi* 'Calf Pond', with the result that these names occur near each other in various regions with suitable cultural conditions. Similar reasons would result in, say, *Myllyjärvi* 'Mill Lake' and *Kirkkojärvi* 'Church Lake' to be found near each other with a surprising regularity.

Likewise, it is natural that similar descriptive elements occur in different places where the terrain is similar. A

muddy and slightly swampy terrain would motivate both *Paskolampi* 'Shit Pond' and *Liejulampi* 'Mud Pond'; in another region, geological processes would result in a terrain that would have both *Kaitajärvi* 'Narrow Lake' and *Hoikkajärvi* 'Thin Lake'

In the case of both cultural and natural connections, the underlying phenomena are often not very easy to analyse. On the other hand, these attraction groups result from causes that are rather clearly outside the linguistic sphere, and in a study of naming patterns it seems reasonable to concentrate on other pairs.

In the "interesting" set of attraction groups it is possible to see two main types of patterns. First of all, there are several pairs that result of inductive naming; second, there is a variety of pairs that are contrastive. Furthermore, both these main patterns would appear to be very productive.

The inductive patterns include cases of naming a smaller lake after a larger one, so that e.g. *Mäntyjärvi* 'Pine Lake' would have *Mäntylampi* 'Pine Pond' nearby. Also, there are several pairs where each of the names has an element specifying either the size or direction: for instance, *Iso Haukilampi* 'Great Pike Pond' and *Pieni Haukilampi* 'Small Pike Pond', or a series of *Alalampi* 'Low Pond' — *Keskilampi* 'Middle Pond' — *Ylilampi* 'High Pond'. As seen from the last example, the distinction between these groups and those resulting from contrastive naming is not always clear.

The contrastive patterns, in turn, usually involve varying some sort of theme. For instance, two lakes can be named after different species of fish, such as *Ahvenlampi* 'Perch Pond' and *Haukilampi* 'Pike Pond', or they can vary a theme that appears to result from the physical characteristics of one of the lakes, like *Mustalampi* 'Black Pond' vs. *Valkealampi* 'White Pond'. In most of the cases only the modifier changes, but there are cases where the modifiers appear to be contrasting while the heads denote other differences in the lakes: for instance, there are such groups as *Valkeajärvi* 'White Lake' — *Mustalampi* 'Black Pond' or even *Valkeinen* 'The White' — *Mustalampi* 'Black Pond'.

Discussion

It is possible to consider the attraction groups as resulting from naming patterns in the sense Šrámek (1972) defines the term. Likewise, this study appears to validate the hypothesis expressed by Pamp (1991) and others: there is a clear general tendency towards analogy even when other motivations exist as well. Incidentally, this also seems to indicate that at least a large number of place names were originally coined as names; this is in contrast with the traditional view, first proposed by Leibniz (1710), that proper names in general originate as appellative constructs. On the other hand, however, it also seems that making the distinction between analogy and these other motivations is not always — perhaps even not often — easy, and similarly the distinction between motivational and structural patterns, or *Ausgangsstellungsmodell* and *wortbildende Modell*, is not always as clear as Šrámek made it sound.

I shall now, therefore, change my point of view to that of Construction Grammar (Fillmore—Kay 1995). One of the main theses of this theory, often left implicit, is that there is no real difference between lexicon and grammar; in terms of onomastics this means that each naming pattern can be expressed as a single construction that explains both its semantic and structural properties.

The main ideas may look familiar to those who are acquainted with Cognitive Grammar. This is to be expected, as the two theories are really functionally equivalent and philosophically compatible with each other; this observation is even made in one of the seminal works of Cognitive Grammar (Langacker 1991: 8). While neither theory has been systematically applied to onomastics or its neighbouring fields, there have been some preliminary studies that indicate its usefulness (e.g. Inglis 2004). For this study I have chosen Construction Grammar, because the notation is better suited for the current work.

Most of the lake names in this study are instances of the rather general construction (1) in figure 2. That is, the names consist of some sort of modifier followed by a head that denotes the type of place.[3] The modifier can be

[3] Labelling the function attribute "modifier" as *syntactic* is somewhat arbitrary, and my choice is based mainly on two considerations. First, in Construction Grammar there is not supposed to be a clear division between syntax and morphology — and in toponyms, especially such as (1.3), the distinction really is not very clear. Second, the question of semantics of proper names is worth much more comprehensive discussion than would be possible in this article, so for

either an adjective, as shown in subtype (1.1), or a noun in either nominative (1.2) or genitive case (1.3). Of these, types (1.1) and (1.3) have appellative homonyms — modulo orthography — while type (1.2) occurs mostly in proper names.

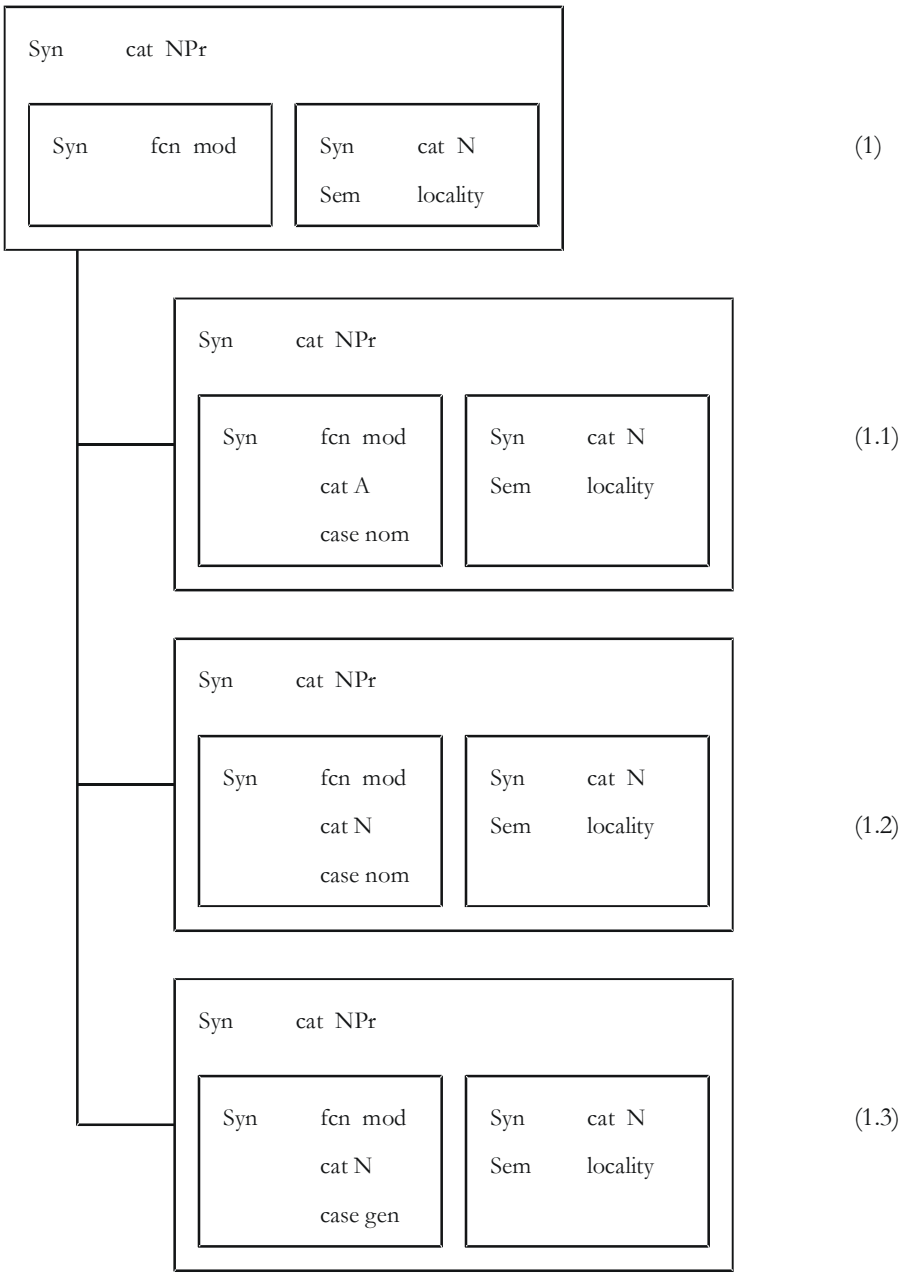


Figure 2: Basic constructions for forming toponyms

There are some cases where the type (1.2) can be used for appellative descriptions, such as *lintujärvi* 'a lake frequented by birds', or possibly even *haukijärvi* 'a lake especially good for fishing pikes'. However, many of the toponyms do not have a meaningful appellative homonym, so e.g. *Housulampi* 'Trouser Lake' typically refers to the shape of the lake instead of any tendency of trousers to be found near the lake. There are some similarities between names of this latter type and appellative bahuvrihi compounds such as *puupää* 'blockhead' (cf. e.g. Malmivaara 2004), but while these phenomena may well turn out to be related they are by no means identical.

Of the subtypes, (1.1) and (1.2) are the most common. In other words, most names where the modifier is a noun

---

the moment I would like to avoid the issue as far as possible. For these reasons I prefer to call the attribute syntactic instead of morphological or semantic.

do not, as such, have an appellative homonym but are immediately identifiable as proper names. Even names that have a corresponding common noun construction occur often as members in attraction groups, which seems to indicate that they — or at least a large fraction of them — did not originate purely as descriptive designators; rather, the naming process included analogy as a major component.

The widespread use of analogy is quite consistent with a constructionist or cognitive view. In fact, Onikki-Rantajääskö (2001: 34—36) suggests that the distinction between analogy and more systematic patterns is not necessarily clear or even relevant. In other words, it is to be expected that people generalise from an expression — in this case an existing name — and use the resulting construction to form new names. Figure 3 shows how this works: the name in (1.2.1) is used to give the construction (1.2'), a rather more strictly stated version of (1.2), above. This is subsequently used to form the name in (1.2.2). Traditionally this method of forming names is called analogy; here the same process is simply interpreted in terms of a human tendency to generalise even minute amounts of data into rules (cf. Langacker 1991: 48).

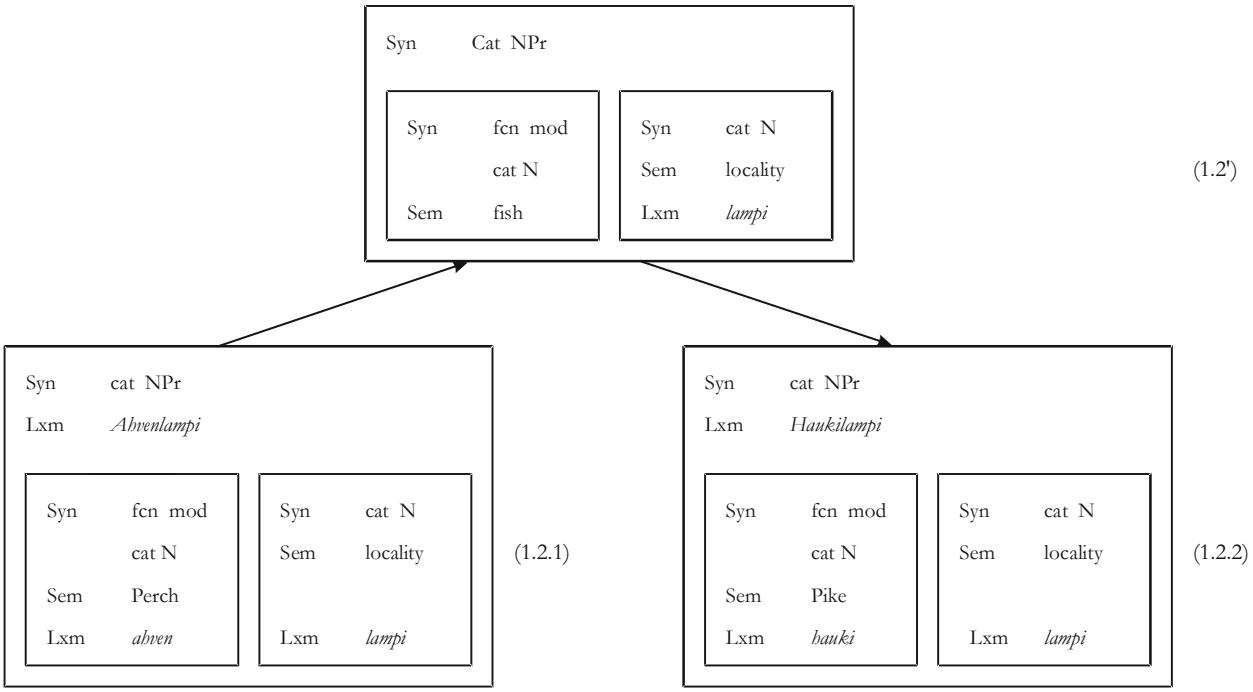


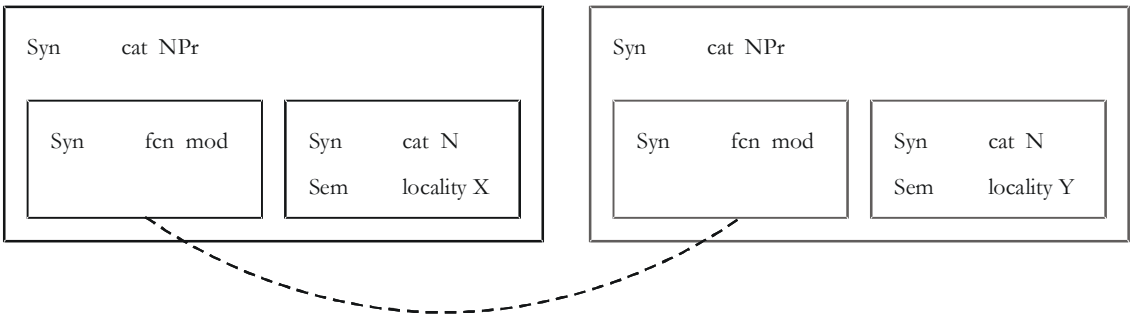
Figure 3: Construction induced from *Ahvenlampi*

Orthodox practitioners of Construction Grammar would not call these ad-hoc generalisations *constructions*, but rather reserve that term for structures that are established parts of the language. Instead, they would use concepts like *coining* (e.g. Fillmore 1997; Kay 2002). However, their distinction between productive constructions and unproductive patterns of coining seems to be too strict for my present needs: the phenomenon I am trying to describe is neither systematically productive nor unproductive, but somewhere between these extremes. On a more general note, this productivity problem, so to say, may not be restricted to toponyms: there are some indications that semi-productive patterns are reasonably common, and one of the challenges to the descriptive apparatus of Construction Grammar is how to represent these. Figure 3 is, in this respect, an interim solution that attempts to duck the issue for now. Still, it is worth noting that generating constructions from established lexical elements would give a certain amount of symmetry to a theory where the opposite is already normal.

With this caveat, the contrastive naming patterns seem reasonably easy to describe in terms of constructions. The case of inductive names is also relatively straightforward, although the situation here is somewhat different. The key point, as I see it, is that these names are derived from existing toponyms and are often meaningful only when one knows that the other name exists in the neighbourhood.[4] In fact, as Nicolaisen (1990) points out, this sort of inductive naming is an essential part of the toponymic system: it makes it possible to increase the number of items in it without having to coin an unlimited number of independent names. The obvious way to describe this is to

[4] This is especially true in cases where the two names refer to different types of places: for instance, the name *Ahvenkorpi* 'Perch Waste' becomes much more understandable when one notices that this slightly swampy patch of forest is near *Ahvenjärvi* 'Perch Lake'.

include the other name in the construction in some manner. Thus, names like *X Lake — X Pond* can be described as in figure 4. Here the names have different heads, denoting the type of place; however, they both share the modifier.



(2)

Figure 4: Construction of *X Lake / X Pond*

An alternative way to view this phenomenon would be to interpret it as a variant of figure 3, but consider the physical neighbourhood as some sort of context. In this approach, the relationship between the two names would be considered as something analogous to connectivity in text. In this interpretation, the key concepts are those of augmentation and grounding, as used in Cognitive Linguistics (Langacker 2001). The construction itself is similar to that in figure 3, but it is augmented to include a reference to the name of a near-by place, even though the form remains unchanged.

Returning to the interpretation that resulted in figure 4, patterns like 'Greater X' vs. 'Lesser X' (and similar pairs involving other semantically contrasting modifiers, such as directions) can be described as in figure 5. Here construction (3) is the general type: a modifier followed by a head which is in itself a proper name. In subtype (3.1), the new name is accompanied by an unmodified name, e.g. *Pieni Haukilampi* 'Small Pike Pond' next to *Haukilampi* 'Pike Pond'. In subtype (3.2) the unmodified name does not exist in itself (or it no longer exists); rather, there are two names that include opposite modifiers, such as *Pieni-Valkeinen* 'Lesser White' next to *Iso-Valkeinen* 'Greater White'.

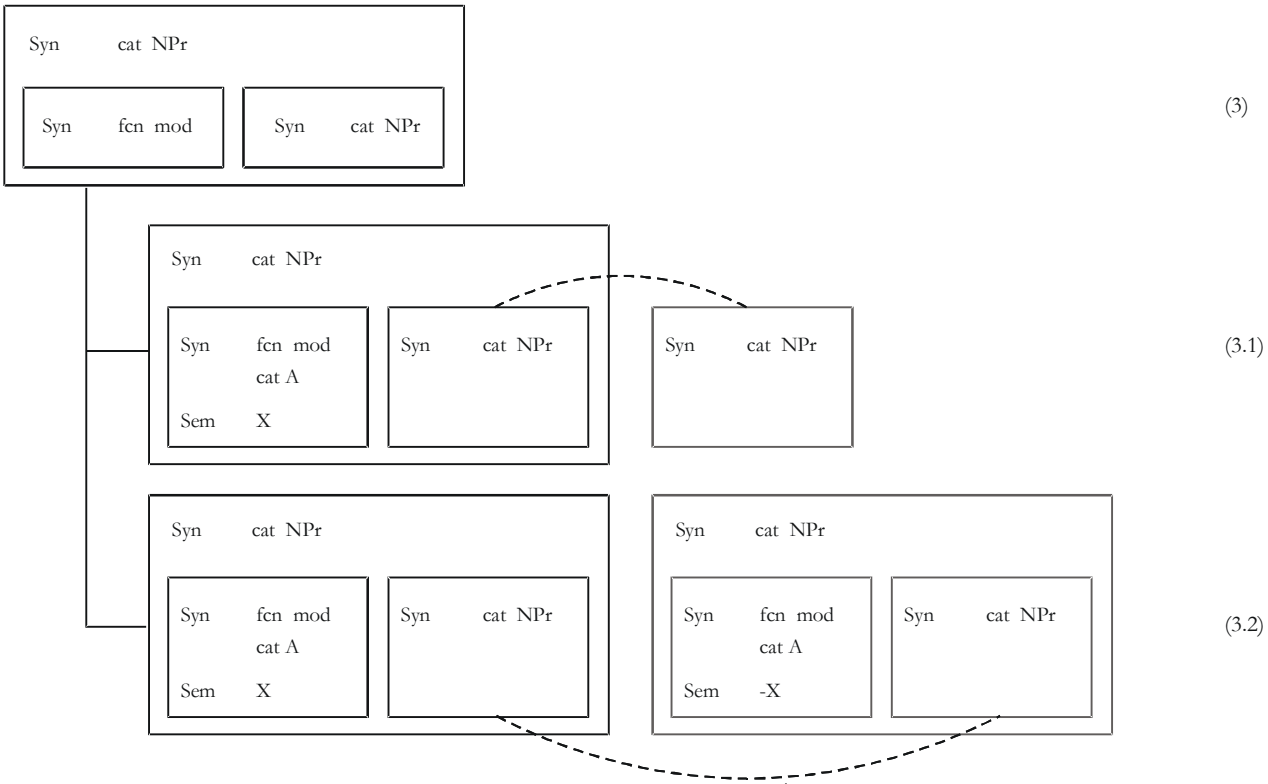


Figure 5: Construction of *Greater / Lesser X*

Conclusions

In summary, it is clear that the old hypothesis was right: in general, analogy plays an important role in naming lakes even in cases where the names are motivated by the features of the place. However, there is more to this issue than that, and it is becoming clear that the question should not be phrased in such an either-or fashion. Rather, analogy only works if the other conditions allow it to be used; and likewise, the selection of descriptive elements is fundamentally influenced by analogy.

Moreover, the traditional term *naming pattern* does not feel right. It tends to imply that these "patterns" are reasonably stable and that they can be clearly defined, but this does not seem to be the case: on the contrary, most of them are rather specific and have obviously been used in a rather ad-hoc fashion. As a consequence, I have a growing conviction that the mechanisms of the phenomena which have been grouped under the umbrella of *naming patterns* are in need of further study. Quite likely it is not just the term which is inadequate, but rather the entire concept of naming patterns needs revision.

To address both of these issues, it would seem that Construction Grammar and Cognitive Grammar provide a suitable framework to describe the interactions between place names. However, there have been few publications on applying either of these theories to anything resembling the toponymic processes, and it is evident that there is a lot of room for further work.

References

Agrawal, Rakesh and Ramakrishnan Srikant 1994: *Fast Algorithms for Mining Association Rules in Large Databases*. In *Proceedings of the Twentieth International Conference on Very Large Data Bases (VLDB'94)*, pages 487—499.

Bailey, Trevor C. and Anthony C. Gatrell 1995: *Interactive Spatial Data Analysis*. Longman Scientific & Technical, Harlow, Essex.

Fillmore, Charles 1997: *Lecture on idiomaticity*. Notes available at <http://www.icsi.berkeley.edu/ kay/bcg/lec02.html>.

Fillmore, Charles and Paul Kay 1995: *Construction Grammar*. CSLI Lecture Notes. Center for the Study of Language and Information, Stanford, CA.

Hand, David J., Heikki Mannila and Padhraic Smyth 2001: *Principles of Data Mining*. MIT Press.

Inglis, Douglas 2004: *Cognitive grammar and lexicography*. In Second Asia International Lexicography Conference. <http://crcl.th.net/sealex/Inglis_CogGramDict.pdf>.

Kay, Paul 2002: *Patterns of coining*. In Second International Conference on Construction Grammar. Electronic version, <http://www.icsi.berkeley.edu/ kay/coining.pdf>.

Kiviniemi, Eero 1977: *Väärät vedet. Tutkimus mallien osuudesta nimenmuodostuksessa*. Suomalaisen Kirjallisuuden Seuran toimituksia 337. SKS, Helsinki.

Langacker, Ronald W. 1991: *Foundations of Cognitive Grammar*, volume II: *Descriptive Application*. Stanford University Press, Stanford, CA.

— 2001: *Discourse in Cognitive Grammar*. Cognitive Linguistics, 12(2), 143—188.

Freiherr von Leibniz, Gottfried Wilhelm 1710: *Brevis designatio meditationum de Originibus Gentium, ductis potissimum ex indicio linguarum*. In *Miscellanea Berolinensia ad incrementum scientiarum, ex scriptis Societati Regiae Scientiarum exhibitis edita*, volume I, pages 1—16. Johan. Christ. Papenius, Berolinum. Electronic facsimile, <http://www.bbaw.de/bibliothek/digital/>.

Leino, Antti, Heikki Mannila and Ritva Liisa Pitkänen 2003: *Rule discovery and probabilistic modeling for onomastic data*. In Nada Lavrač, Dragan Gamberger, Ljupčo Todorovski and Hendrik Blockeel (editors), *Knowledge Discovery in Databases: PKDD 2003*, Lecture Notes in Artificial Intelligence 2838, pages 291—302. Springer, Berlin.

Leskinen, Teemu 2002: *The geographic names register of the National Land Survey of Finland*. In Eighth United Nations Conference on the Standardization of Geographical Names.

Malmivaara, Terhi 2004: *Luupää, puupää, puusilmä. Näkymiä sananmuodostuksen analogisuuteen ja bahuvrihiyhdyssanojen olemukseen* (Perspectives on analogy in word formation and the essence of bahuvrihi compounds). Virittäjä, 108(3), 347—363.

Mannila, Heikki, Hannu Toivonen and A. Inkeri Verkamo 1994: *Efficient algorithms for discovering association rules*. In Usama M. Fayyad and Ramasamy Uthurusamy (editors), *Knowledge Discovery in Databases, Papers from the 1994 AAAI Workshop (KDD 94),* pages 181—192. AAAI Press, Menlo Park, CA.

Nicolaisen, Wilhelm F. H. 1990: *The growth of name systems*. In Eeva Maria Närhi (editor), *Proceedings of the XVIIth International Congress of Onomastic Sciences*, volume 2, pages 203—210. Helsinki.

Onikki-Rantajääskö, Tiina (2001). *Sarjoja. Nykysuomen paikallissijaiset olotilanilmaukset kielen analogisuuden ilmentäjinä*. Suomalaisen Kirjallisuuden Seuran toimituksia 817. SKS, Helsinki.

Pamp, Bengt 1991: *Onomastisk analogi*. In Gordon Albøge, Eva Villarsen Meldgaard, and Lis Weise (editors), *Tiende nordiske navneforskerkongres, Brandbjerg 20. 24. maj 1989*, Norna-rapporter 45, pages 157—174.

Ripley, Brian D. 1976: *The second-order analysis of stationary point processes*. Journal of Applied Probability, 13, 255—266.

Šrámek, Rudolf 1972: *Zum Begriff "Modell" und "System" in der Toponomastik*. Onoma, pages 55—75.