Toponymy and language

Tjeerd Tichelaar

Toponymy is the science that has as its subject the study of geographical names or toponyms[i]. As all other names, toponyms belong to languages. Names in general are only rarely randomly chosen, and this is especially true in the case of geographical names. Whether they carry a physical meaning like Mont Blanc ('White Mountain'), or they were coined to honour someone (Washington, District of Columbia), to commemorate some historic event or to make clear to whom the named object belonged (Paris, from Latin 'Lutetia Parisiorum' = 'Lutetia of the [Gallic tribe named the] Parisians'), in all cases they once used the vocabulary and followed the grammatical and orthographic rules of a certain language.

Languages are the subjects of the science called linguistics. Therefore, anyone handling geographical names needs to have some basic linguistic knowledge, both in general terms and specifically pertaining to the language situation of the area of survey.

Toponymy and linguistics

People from different professional backgrounds may be allured to some kind of study of geographical names. To linguists specializing either in the historical or genealogical aspects of specific languages, or in the taxonomy of languages in general, toponyms contain a treasure of ancient language elements which allows them to under build their theories or test their hypotheses. Likewise, historians may use toponym research to reveal ancient movements of peoples, or get a hint of cultural exchange patterns in forgotten ages. Moreover, recurrent name elements are known to store information on the history of settlement and land reclamation, the economic activities of the original settlers, and political developments.

Topographers and cartographers often bear a less theoretical interest in toponymy: they simply need to know by what name(s) every object to be mapped has to be known and recorded.

As far as the last mentioned category of professionals does not study geographical names for the sake of the names themselves, but rather wants to constitute a set of rules, or *standards*, defining what should be considered 'right' and 'wrong' in the cartographic naming practice, they are involved in what we call *applied toponymy*. Even if exhaustive linguistic knowledge is not required to be able to practise this specific kind of applied toponymy, a basic understanding of the linguistic and historic context of the geographical names within the area of study is certainly indispensable.

The relation between names and language

At the moment a name is given to an object, the language of the name-giver provides both the elements needed and the structure to join them together. The elements consist of semantic and morphologic units – units of meaning and form - called *words* and *morphemes*. The former are the smallest units that may occur independently, the latter the even smaller particles, like suffixes and affixes forming part of or joined to them. The structure is provided in the form of a set of rules called *grammar*, that defines the way the language can be used to convey (*communicate*) meaning. An important constituent of grammar is the *syntax*, determining the way words should be linked together into larger semantic conglomerates. Most names start their existence as such a semantic conglomerate.

The linguistic abracadabra above may easily be clarified by picking the first name that comes to mind, for instance: Stratford-upon-Avon. This English town that became world famous thanks to the birth of William Shakespeare clearly consists of three elements, which are, obviously in accordance with some syntactic rule specifically applying to English names, separated by hyphens. Two of the elements start with a capital, the one in the middle doesn't: again a syntactic rule. As a capital initial letter is commonly used in (Roman) written language to denote that a word is either the beginning of a sentence or a name, we get the idea that both 'Stratford' and 'Avon' are names in themselves, and 'upon' is not. We need to know that 'upon' is a preposition, meant to establish a situational link between 'Stratford' and 'Avon'. Both of the remaining elements of this name also enclose a meaning for themselves, that at the time of the name-giving must have been considered important: this meaning had to ensure that upon mentioning it would make clear which geographical object was meant, without anyone needing to point at it.

'Stratford' appears to be an Anglo-Saxon (Old English) name, consisting again of two semantic units, namely *strat* and *ford*. 'Strat' is adopted from the Latin word *strata*, meaning 'paved road'. It was a Latin (Roman) name for something the Romans made and left behind for the Anglo-Saxons, who did not know it themselves. The paved road referred to was in this case the Roman road from Alcester (ancient Alauna) to Tiddington, both of them Celtic settlements fortified by the Romans. 'Ford' is an Anglo-Saxon word that still exists in modern English, meaning 'part of a river shallow enough for people to cross it'. So 'Stratford' was obviously the place where one would cross the river when following the Roman road. If one would mention this, anyone would know which site was meant without someone having to go there and point at it.
The addition 'upon Avon' became obviously necessary when the place became important enough to be mentioned to people who might also know other places where Roman roads crossed rivers, or other towns named 'Stratford'; to these people the mentioning of 'Stratford' alone might not provide enough information. The name 'Avon' itself is Celtic; it simply means 'river'; it is still the name of the river flowing through Stratford-upon-Avon.

It is thus clear that the name of this town really started as a 'semantic conglomerate', even though today it's meaning to most people is just 'Shakespeare's birthplace'. The Roman road became forgotten, its asphalt successor being not special enough to be mentioned, and the ford lost its importance once bridges were built. The addition 'upon Avon' remained worth mentioning because of the existence of another Stratford, namely in the Greater London conurbation.

Languages of the world

The languages currently spoken and written in our world are as diverse as the societies making use of them. Although there are different conceptions of what should be considered a *language* and what is merely to be seen as a local or regional variety of speech or *dialect*, a contemporary count carried out by the Summer Institute of Linguistics results in a total number of nearly 7,000 languages[ii]. This stunning number has by now been classified into a hundred 'families' and 'phyla', while 96 languages still remain unclassified and 30 others are considered to be 'language isolates', meaning that they are not related to any other known language. The *family* metaphor is used for any grouping of languages that is thought to have diverged from one common ancestor, whether this is supported by real historic evidence or linguistic analysis itself provides indications in this direction. The term *phylum* is applied to

groupings of languages where such evidence or indications are missing, but nevertheless some kind of relationship is suspected. Amongst linguistic researchers there is a quest for yet unnoticed kinship ties between separate families, and progress of language-genealogical research generally leads to a reduction of the number of unrelated families and language isolates.

Toponymic importance of individual languages

From a global point of view, obviously not all language families are as important, as far as numbers measure importance. More than 75% of all languages belong to only 10 of the 100 recognized families, while judged by the numbers of speakers, two-thirds of the world population speak languages belonging to only two families (Indo-European and Sino-Tibetan).
To the topographic-cartographic toponymist, however, other numbers may be even more relevant: after all, the number of geographic names to be dealt with is not so much dependent on current numbers of speakers, as it is on the geographic extent of the area to be surveyed and the scale of mapping the survey is carried out for. Topographic map series of a certain scale use to cover a complete country, irrespective of differences in population density.

The implication of this last observation may be quickly illustrated by the following real-world example. The 27,000 Nenets-speaking Samoyeds are mainly (former) nomadic reindeer herdsmen. For many centuries their ancestors dominated a huge area in northern Siberia, in which they named all terrain features (streams, hills and so on) that had any meaning to them. Despite the small number of Nenets speakers - maybe equivalent to the number of inhabitants of just a few apartment blocks in New York City - their language has to be taken into account in an area of four or five times the size of the United Kingdom.

To further illustrate the complexity of defining the importance of a language to toponymy, let us stick to the Siberian Nenets a little longer. Especially during the last century, the Nenets homeland has received an influx of Russian settlers, who soon outnumbered the Nenets in their own provinces – be it that the newcomers settled in just a few urban settlements. Furthermore, from the 1950-s on the Soviet authorities where rather successful in putting the nomadic lifestyle of the Nenets to an end[iii]. But the many streams, lakes and other physical landscape elements had for long been named by then. And most of them had been named only once: by the Nenets. By the nature of their nomadic lifestyle, in contrary to that of the new urban settlers, to them little in the landscape was without meaning and therefore without the need of a name.

Two other issues have to be taken into account when evaluating the relationship between numbers of speakers and the importance of a language from a toponymic point of view. One involves the level of geographic (or regional) attachment of a language, the other the ‘historic rights’ the speakers of a language enjoy to the land where they settled. Both can be easily illustrated reviewing the situation in one of the most stable countries we can think of.

Although we like to think that in Sweden people speak Swedish, the Summer Institute of Linguistics lists 14 main languages for this country – excluding a deaf sign language – as well as 19 other languages used by immigrant communities. To start with the main languages, Swedish has indeed by far the largest number of speakers, in 1986 amounting to 93% of the country’s just over 8 million inhabitants. Next in numeric importance ranks

Skånska, with 1,5 million speakers, followed by Finnish (200,000), Tornedalen Finnish (60,000 to 80,000), Jamska (30,000), Tavringer Romani (25,000), Northern Saami (4,000}, (2,500), Lule Saami (1,500), Kalo Finnish Romani (1,000 to 2,000), Vlax Romani (1,500), Dalecarlian (1,500), Southern Saami (300), Pite Saami (50), and Ume Saami (50). Skånska and Dalecarlian are regional East-Scandinavian languages (just like Swedish), Jamska is West-Scandinavian (like Norwegian); all are lacking any official status, although linguists consider them different languages. Their speakers are almost all bilingual in Swedish.
People formerly known under the derogatory name 'Lapps' speak the Saami languages; they belong to two possibly unrelated subgroups of the Uralic language family. They may have settled in Sweden a millennium or so earlier then the Indo-European Swedes, and were by the beginning of the Christian era at least the sole inhabitants of the northernmost one-third of the country. So, for toponymic purposes, the geographic range of their languages is very much larger then their current area of settlement – which is, in the sparsely populated north of the country, still extensive enough.

Next to the Saami languages, three gypsy languages are mentioned: the two (Indo-Aryan) Romani languages and the Germanic 'Tavringer Romani'. Both Kalo Finnish Romani and Tavringer Romani are spoken by descendents of gypsies deported from Scotland in the early 16th century; the ancestors of the Vlax Romani speaking gypsies (Lovari and Kalderash) arrived more recently, escaping 500 years of slavery in Romania.
While both Saami and Romani may be considered semi-nomadic, the Saami do have much stronger ties to the area they live in – it is where their culture and language developed – than the gypsies. They arrived much earlier, but – more importantly – they were for long the majority population or even the sole inhabitants of their Swedish homeland, while the gypsies were always only a small minority, even on a the smallest regional level. Their presence in Sweden nevertheless antedates the presence of Germanic Swedes in parts of the Saami homeland, so there may be discussion about who may be considered 'native' and who may not: from the point of view of the Saami, both Germanic Swedes and gypsies might be considered Indo-European intruders.

The gypsy immigrants never colonized empty spaces, but were, because of their non-primary economic specialization, attracted by already existing native societies. This is why they were not granted rights, or at least had no opportunity, to give their own names to yet unnamed geographical objects. The same holds for more recent immigration communities that entered Sweden either as regular immigrants, as imported labour, or as refugees from abroad. In spite of their numbers – the 120,000 Servo-Croats, 50,000 Greeks and 35,000 Spanish for instance quite spectacularly outnumber the Saami communities – they lack the geographic attachment and recognized 'historic rights' the Saami clearly possess. And so their meaning for toponymy is likely to be negligible. The same goes for the 200,000 speakers of Finnish, who are all 1st to 3rd generation immigrants. The Tornedalen Finnish, however, are native inhabitants of parts of the county of Norrbotten, their ancestors having settled there (in Saami territory) in the 12th century, which is earlier than the Swedes. Their language is only partly mutual intelligible with standard Finnish.

Toponymic importance of linguistic status

Just like in the context of toponymy numbers of speakers do have another weight than they have from a general linguistic point of view, the question whether or not a language is being officially recognized as such also has less importance in a toponymic sense. This so-called

question of linguistic status – is a specific system of common speech to be considered a real 'independent' language, or, alternatively, 'just a dialect' of a 'real' language? – is answered differently depending on the considerations of who is asked. Political considerations in this may prove to be dominant above any linguistic criteria.

Toponymic terminology includes a couple of status qualifiers. An *official language* is 'a language expressly adopted by the government of a country … and employed as a language of administration'. A *non-official language* is 'a language that lacks official status in a particular legally constituted entity'. A *dialect* is 'a variety of language which is distinguished by phonological and/or morphological characteristics that give it a distinctive identity'. A *literary language* is a 'written form of language regarded as the desirable standard for works of literature'. A *national language* is a 'language in widespread current use throughout a given country or in part of its territory …', and it '… may have or may not have the status of an official language'. A *minority language* is 'any language not used by a significantly large part of the country's population'. A *principal language* is 'in a linguistic community where more than one language is in use, that language which has greatest currency'. A *living* resp. *dead language* is 'any language spoken today, resp. not longer spoken'.

Language vs. dialect

Especially the status distinction between *language* (real) and *dialect* (just a variety of a language) is treated differently by politicians and linguists. The issue of language vs. dialect is known to lead to emotional debates amongst political groupings – at times even culminating in inter-regional conflicts – and scientific hair-splitting amongst linguists. Vernaculars that show too little discrimination to be considered different languages by analytical linguists – like Romanian and Moldavian, Serb and Croat, or Bulgarian and Macedonian – may nevertheless be officially defined as such because of political compulsion, while linguistically distinctive non-official languages like Lower Saxon in Germany and the Netherlands (formerly influential as the language of the Hansa Federation), Skånska, Jamska and Dalecarlian in Sweden (referred to earlier), or any of the regional languages in Italy, are often disposed of as 'dialects' of the official languages (German, Dutch, Polish, Italian) in the respective countries where they are spoken. Alternatively, the 1.3 billion ethnic Chinese citizens of the People's Republic of China stress their unity by considering the Sinitic languages they speak, at least eight of which are mutually unintelligible when spoken, as one single Chinese language, thereby making use of the unique circumstance of sharing an ideographic script that ensures at least mutual intelligibility of the written language.
Political developments cause promotion of 'dialects' to 'languages' and vice versa, even if linguistically nothing changes.

In comparative linguistics, the status of *language*, *dialect* and *sub-dialect* may be awarded to a vernacular attached to the same branch of the genealogical tree on the ground of lexical correspondence – the percentage of shared vocabulary – and grammatical similarity. In the classification of Austronesian and Papuan languages published in Wurm and Hattori's Language Atlas of the Pacific Area, for instance, languages of the same family generally share between 20% and 80% of their basic vocabulary, while for dialects of one (theoretically defined) language this percentage is over 80 and the grammar must be near-identical. Sub-dialects of one (again theoretically defined) dialect must even be more similar to each other[iv]. Of course, the languages, dialects and sub-dialects that are the subject of these classificatory efforts, the units that we will generically refer to here as *vernaculars*, are identified as such

(and named!) because of their *de facto* existence as systems of communication belonging to a certain distinctive community - be it a tribe, a village, a 'nation' or whatever.

From a linguistic point of view, a standard language and regional languages commonly considered dialects may be attached parallel to each other to the same branch of the family tree. In this case, the standard language usually developed as a codified form of one of the dialects in a *dialect chain*, to be seen as a continuum covering a certain geographic area, where local vernaculars and dialects gradually flow into each other. Immediately contingent vernaculars are then quite similar and mutually intelligible, whereas the vernaculars on both ends of the chain show a maximal difference and lack of mutual intelligibility. The codification of the vernacular used in a certain sub-area within the range of this dialect chain into the standard language, subsequently gaining official status, is than a matter of historic coincidence. For instance, one of the branches of the Romance section of the Indo-European languages ends in an Italian dialect chain, that is usually 'cut' into dialect segments mostly carrying the names of historic regions: Tuscan (Tuscany), Umbrian (Umbria), Laziale (Latium), Central Marchigiano (Marche), Cicolano-Reatino-Aquilano (distinctive local vernaculars in the central Apennine border area of Latium, Marche and Abruzzo), Abruzzese (Abruzzo), Molisano (Molise), and Pugliese (Apulia). All of these dialects developed in and after the Roman era out of the Latin language used there in those days. Because Dante Alighieri was, in the late 13th/early 14th century, the first to use the popular (degenerated) form of Latin spoken in his native area, Tuscany, to write influential literature, it was out of his Tuscan dialect that the standard 'Italian language' was born. For the time being, the other dialects did not suffer under the dominance of this new literary standard, because Italy was still divided into a large number of different states and foreign possessions. When in 1861 the country chose to become unified – as part of a 'nation-building' process in which linguistic affinity among at least the ruling elites did play an important role – only 2.5% of the population mastered the standard language.

Fortunately toponymists define a language as 'a system providing a means by which the members of a community can communicate orally and/or graphically', i.e. without respect to its status as any class of language or dialect. To toponymy, a dialect may be worth as much as or even more than an official language, as the majority of toponyms has once been created by the (local) community. Besides, even if the vocabulary of a (sub-) dialect is for 95% similar to that belonging to the dialect or language it is supposed to be a variation of, the 5% difference will most likely include exactly those terms often occurring in geographic names, namely those terms traditionally close to the communities' daily experience (generic terms like *water*, *river*, *lake*, *forest*, *village*; adjectives of colour, size etc.).

Official language

It is worth noticing that in large parts of the world, the official language is not even the language actually spoken today in the largest part of the country. Especially where the official language is a foreign language (for instance anglophone and francophone Africa) or a relatively new language developed from a *lingua franca* (for instance Indonesia), the official language is the language of just a few toponyms, or even none at all. Here the toponyms belong to the local vernacular, whether this is considered a language, a dialect, or a sub-dialect. The toponymist must then be accompanied by linguists having knowledge of these local or regional vernaculars, in order to be able to correctly interpret the names, as well as correctly define their graphic representation (writing).

Dead and disappeared languages

Because toponyms generally (although not always) outlive their creators, locally disappeared and even 'dead' languages are not per definition deprived of their importance from a toponymic point of view. Dead languages often leave their traces both in the vocabulary of their living successor languages and, much more so, in geographical names. This is a well-known fact to historical linguists, who make indeed grateful use of toponyms in their efforts to reconstruct so-called proto-languages (disappeared common ancestors of modern languages belonging to the same family), as well as trace *substrates*, residues of local predecessor languages in unrelated immigrant successor languages. Especially hydronyms (water names) have a reputation of being very ancient, and, for instance, antedating the 4th and 3rd Millennium B.C. Indo-European immigration into Europe. These substrates are held responsible for a major part of the diversification between the branches of the Indo-European language family; the vocabulary of the Germanic languages, for instance, is thought to contain a large number of pre-Indo-European words, maybe inherited from the thriving 4th Millennium society that build the numerous tumuli and megalithic monuments in north-western Europe. Also the Greek geographical generic term meaning 'sea', *thalassos*, is supposed to be of pre-Hellenic and pre-Indo-European ('Pelasgian' descent) – suggesting that this famous seafaring people was not yet so familiar with the sea at the time it reached its present homeland..

A quick survey of the geographical names in a well-known country like the United Kingdom will further illustrate the arguments expounded above.

The official language of the United Kingdom is English. Besides English, the dwindling Celtic languages Welsh and (Scottish) Gaelic also have official status on a sub-national level. English is a Germanic language, which developed from the closely related languages of Anglian and Saxon immigrants in the 5th century A.D. In the part of the kingdom currently called England, Anglo-Saxon and Jutish invaders, earlier than their Germanic language(s), superseded a mixed Roman and Brythonic Celtic aristocracy ruling a partly Romanised, but largely still Celtic (Brythonic) speaking population. The part of the Brythonic population most strongly opposing assimilation with the Anglo-Saxon language and culture fled the Germanic invaders to take refuge in present-day Wales, the border area of England and Scotland (Cumbria and Strathclyde) the south-western peninsula of England, and the peninsula of Brittany in continental Gaul - currently France. In Scotland, at the same time, a Pictish population speaking an as yet unknown language that had taken refuge there for the Roman invaders of the island, four centuries earlier, were gradually superseded by so-called Goidelic Celts (Gaels, *Scoti*) invading their homeland by sea from Ireland. The Brythonic and Gaelic newcomers in Scotland were, although both Celtic, distinctive enough not to understand each other's language.
Starting from the 8th century, new Germanic immigrants invaded the country: Norwegian and Danish Vikings took possession of and effectively colonized large parts of both Scotland and England, to be eventually (in the 11th century) expelled again by the Anglo-Saxons. Even before the last Norwegians were ousted, however, Anglo-Saxon dominance itself came to an end by an invasion of yet another Viking aristocracy: this time the already Romanised (French-speaking) Normans successfully claiming the English throne.

Although at present the English language is, apart from being the only nation-wide official language, the mother tongue of more than 99% of the native inhabitants of the United Kingdom. But before Anglo-Saxon or English became dominant, Pictish, Brythonic Celtic, and Latin were for centuries the languages of both aristocracy and (part of) the common people, as were Gaelic Celtic, Norwegian, Danish and French (the latter mostly of aristocracy) after the introduction of Anglo-Saxon. The imprint of some of these languages on the geographical names of the British Isles is at least as large as Anglo-Saxon/English: the large majority of names in Scotland is of Gaelic origin, except in Strathclyde, where many names are either Brythonic or Anglo-Saxon, and in the Northern and Western isles (Shetland, Orkney and the Hebrides), where almost all names are of Norwegian descent; the islands were Norwegian from the 8th until the 15th century, which was long enough for a new Scandinavian language to develop there (Norn, spoken in Orkney until the 18th century). The northern and eastern parts of England show a mixture of Danish – for instance names on *–by* (= 'farmstead, village') - and Anglo-Saxon, while the southeast is predominantly Anglo-Saxon; in the southwest the Brythonic element is dominant. All through England a Brythonic substrate is eminent, as are remains of Latin like the formerly generic elements *caster* or *chester* (Lancaster, Manchester - from *castra* = 'fortress') and –port (from *portus* = 'harbour' or *porta* = 'gate'). Wales is almost completely Brythonic; the Anglicised forms of Brythonic (Welsh) names were with the recent emancipation of the Welsh language returned to their original state, and English names reverted to their Welsh counterparts. In Cornwall in Southwest-England, the Cornish (Brythonic) language, actually extinct (a 'dead language') since 1777, is presently being revived and granted official status next to English on a local level: some Cornish place-names are being restored correspondingly.

The English language itself lost much of its original Anglo-Saxon character because of all subsequent invasions, causing the grammatical structure to be simplified and the vocabulary augmented with a large amount of Scandinavian and French words. Geographical names where in writing often adapted to the language passing by – a nice example is York, going back on a Brythonic personal name Eburos (meaning 'yew man'), maybe the owner of an estate with yew trees where the Romans built their fortress Eburacum; the Anglo-Saxons, ignorant of this meaning, transformed the name through etymological misinterpretation ('popular etymology') into Eoforwic, meaning 'wild boar settlement'. The Vikings taking over the place from the Anglo-Saxons contracted the first part of the name, without bothering for a possible meaning that they didn't understand anyway, into 'Hjor', while they thought to understand the second part as the similar sounding Norse generic 'vík', meaning 'bay' (not very appropriate for the inland town). They were the last to bother at all: the Anglo-Saxons ousting the Danes, just before they themselves had to accept francophone Norman rule, left the name as it was remodelled by the Vikings: Hjorvík; the Anglo-Saxon tongue would ultimately erode this into what it is now: York.[v]

The process of subsequent transformations of names illustrated by the case of York above shows the significance of 'dead' as much as 'living' languages to the development of geographical names. The 'erosion' ultimately yielding the present form of the name does not follow a random path, but is dependent on the phonological characteristics of the 'new' language (the Anglo-Saxon dialect of Yorkshire) as compared to those of the 'old' language (mediaeval Norse or Danish); the regional settlement history, as it also culminates in the local dialect, is decisive. Latin *castra* thus used to evolve into *caster* in the areas of Northern England staying for long out of the grip of the Anglo-Saxons, but tended to become *chester* or *cester* in the more thoroughly anglicised parts of the country. It is thus the phonology of the dialect, not the official language, that determines the ultimate form of the name.

Classification of languages

The following are the most important language families and phyla commonly distinguished, presented here in a more or less geographical order.

1. The Khoisan family

An ancient family of a few scores of languages currently still spoken in the Kalahari and Namib Desert areas of southern Africa, as well as in some isolated areas in Tanzania. The speakers of the Khoi-Khoin (or 'Hottentot') and San (or 'Bushmen') languages are anthropologically unrelated to their African neighbours. The adoption of the 'click'-sounds characteristic to this family by neighbouring Bantu languages of the Niger-Congo family (q.v.) may demonstrate that Khoisan languages in earlier days were native to a larger area than they are at present. The Nama language of Namibia has the largest number of speakers (150,000), followed by Sandawe in Tanzania (70,000); most of the languages are used by a few hundred to a few thousand speakers only.

2. The Niger-Congo family

The most prominent language family of sub-Saharan Africa is the Niger-Congo family. Up to almost 1,500 separate languages are distinguished, belonging to a large number of sub-family level groupings. The exact hierarchical subdivision within this family is still under investigation, and different opinions compete with each other. However, the hypothetical common 'proto-Niger-Congo' ancestor is thought to have ceased to exist as early as 5,000 years ago. The so-called Bantu languages make up the largest of the sub-families, native to Central and the largest part of Southern Africa. Most numerous are Swahili with 5 million first language speakers in the East African countries, an additional 30 million using it as a second language *lingua franca*; Yoruba (20 million) in Nigeria and the eastern part of West Africa; Igbo (17 million) of Nigeria; Fulani (13 million including second language speakers) in West Africa, Wolof (2.7 million in Senegal, an additional 7 million second language speakers); Zulu (9.5 million) of South Africa, Lesotho, Swaziland and Mozambique; Rwanda (9.5 million) of Rwanda and adjacent countries; Lingala (8.5 million including second language speakers) of the Democratic Republic of the Congo, Xhosa (7 million) of South Africa and Lesotho; Shona (7 million) of Zimbabwe, Mozambique, Zambia and Malawi; and Akan (7 million) of Ghana.

3. The Nilo-Saharan family

The 200 Nilo-Saharan languages are spoken in areas on the southern fringe of the Sahara and around the upper course of the Nile, in a zone stretching from Mali in the west to Eritrea and Tanzania in the east. The subdivision of the Nilo-Saharan family is under fierce debate.

4. The Afro-Asiatic family

The Afro-Asiatic family, formerly known as Hamito-Semitic, contains almost 400 languages spoken in Northern Africa and Southwest Asia. In Asia all Afro-Asiatic languages belong to the *Semitic* subfamily, containing amongst others Arabic, Hebrew, Neo-Aramaic languages (descendents of the Aramaic spoken in Roman Syria and Palestine), and a dozen important Ethiopian languages, some of which are official in Ethiopia and Eritrea (Amharic, Tigre, Tigrinya). In Africa, the languages formerly labelled 'Hamitic' are nowadays subdivided into a number of separate sub-families, like *Berber* (in the north-western part of Africa), *Chadic* (in Chad and Nigeria), *Cushitic* (the north-eastern part of Africa from Sudan to Tanzania), *Egyptian* (sole surviving member: the Coptic of the Egyptian orthodox church), and *Omotic* (in Ethiopia).

5. The Indo-European family

Presumably started as a single language of a nomadic cattle-raising people living in the plains of Central Asia, to migrate subsequently to what is now Ukraine and Southern Russia, Indo-European languages spread and expanded since the 3rd Millennium B.C. all over Europe and well into Southwest and South-Asia. The different branches of this family – most importantly *Indo-Iranian*, *Italic* (Latin and the Romance languages born out of it), *Germanic*, *Slavic*, *Celtic* and *Hellenic* – contain the official languages of India, Pakistan, Iran, Bangladesh, Sri Lanka, Nepal, Afghanistan, Tajikistan and Armenia in Asia and almost all the countries of Europe as well as many of their former colonies in other continents. The other branches still represented today are *Albanian*, *Armenian*, and *Baltic* (Lithuanian and Latvian).

6. The Dravidian family

Dravidian languages, some of which enjoy many millions of speakers (Telugu 75 million, Tamil 70 million, Kannada 45 million, Malayalam 35 million), are mainly spoken in the southern part of South Asia. They may have been native to the whole Indian subcontinent by the time the Indo-European languages spread into the Indus and Ganges valleys in the late 3rd Millennium B.C. Indian immigrants also brought the languages to various other parts of Asia, Oceania, Africa and the Americas, but there they did not spread beyond the distinguishable immigrant populations that brought them there.

7. The Caucasian family

Caucasian languages are spoken in a very limited geographical area, roughly coinciding with the Caucasus and Little Caucasus mountains and the valleys in between these mountain ranges. In spite of their modest dissemination, they include the official language of Georgia as well as a large enough number of recognized national languages in autonomous republics within the Russian federation. Whether or not the two major branches of the Caucasian family, *North Caucasian* and *South Caucasian*, are actually related and rightfully included in one single family, is still a subject of linguistic debate.

8. The Uralic family

In Northern Europe, Hungary and Western Siberia, a number of (until quite recently) nomadic peoples and some sedentary ones – most notably the Hungarians, Fins and Estonians – speak languages belonging to the Uralic language family. In spite of the relatively small numbers of speakers (only Hungarian with 14million, Finnish with 6 million, and Estonian with 1.1 million, all official languages in their respective countries, exceed the one million), the languages are important over a territory of millions of square miles. The languages developed from a presumed common proto-Uralic ancestor from as early as the 6th Millennium B.C., and were in an early stage split between a *Finno-Ugric* and a *Samoyedic* branch. The Finno-Ugric branch itself fell apart many centuries ago into a *Finno-Permic* and an *Ugric* branch – this happened so long ago, actually, that it is a difficult job to find but one shared word root in the vocabularies of the leading languages of both branches: Finnish and Hungarian. The Finno-Permic branch is nowadays represented in the Finnish-Estonian area and in the sparsely populated taiga lands to the east of it, up to the western foothills of the Ural mountains. The Ugric peoples fell, as a result of migrations in the first half of the first Millennium A.D., further apart into two from a geographical as well as a cultural and economic point of view opposite groupings: the 14 million Hungarians in the center of the European heartland, and the Ob-Ugric Khanty (22,000) and Mansi (8,000) in the basin of the central and lower Ob River.

9. The Altaic family

Through several waves of migration during the 1st Millennium A.D., the Altaic languages spread from (probably) an area to the east or northeast of Central Asia in a western an south-western direction. All three branches of this family, *Mongolian*, *Tungusic* and *Turkic*, contain languages of famous conquering peoples: Mongolians, Manchus, and different Turkish peoples (Tatars, Turks) respectively. Most prominent nowadays is the Turkic branch, including the official languages of Turkey, Azerbeidjan, Kazakhstan, Uzbekistan, Turkmenistan and Kyrgyzstan, as well as numerous national languages in Russia, Western China and parts of Southwest Asia. Mongolian languages are, except in Mongolia, also spoken in Russia (Kalmyk west of the Caspian Sea, Buryat in the Baikal area), and Northern China. Manchu, once the language of a ruling dynasty in China and for some centuries a *lingua franca* between China and the West, is reportedly still spoken by just a handful of people over 70 years of age – although millions still belong to the official Manchu nationality – but other languages of the Tungusic branch are still alive in vast areas of Eastern Siberia.

10. The Sino-Tibetan family

The fact that more people in this world speak a Sino-Tibetan language than a language belonging to any other family is accounted for by one language only: Chinese, more precisely Mandarin Chinese (900 million speakers). Together with 13 other Chinese (or Sinitic) languages, it even counts 1.3 billion speakers uniquely sharing one and the same written language. The *Chinese* languages actually constitute just one of the branches of this language family, that also includes the *Tibeto-Burman* languages spoken in Myanmar (the official language Burmese and many others), Tibet, and Himalayan areas of Nepal, India and Indo-China. Although Chinese emigrants brought their language with them all over the world, it never spread to non-ethnic Chinese populations elsewhere.

11. The Austro-Asiatic family

In the area in between China and the Malay Archipelago, stretching from Eastern India to the interior of the Malaccan peninsula – actually also including the south-eastern part of China – languages are spoken that are assigned to the Austro-Asiatic language family. The two main components constituting this family are the *Mon-Khmer* branch of Indo-China (extending into eastern India), and the *Munda* languages of India. Mon-Khmer languages include the official language of Cambodia (Khmer, over 7 million speakers), and, although this is being disputed, Vietnam (Vietnamese, 70 million speakers). The Mon language, currently spoken by 1 million people mainly in Myanmar, is an ancient literary language that in the past belonged to an empire stretching well into Thailand.

12. The Daic family

The 70 languages of the Daic, or Tai-Kadai family are spoken in an area to the east of the Austro-Asiatic languages in Indo-China. They include two official languages: Thai (the mother tongue of up to 25 million people in Thailand) and Lao (spoken by 4 million people in Laos, of which it is the official language). Attempts are being made to link the Tai languages to either the Sino-Tibetan or the Austronesian families.

13. The Austronesian family

It is the Austronesian family, and not the Indo-European or the Afro-Asiatic that has the vastest geographical extent as far as native speakers are concerned. Austronesian languages are spoken on islands from the Western Indian Ocean (Madagascar) to the Eastern Pacific (Easter Island, belonging to Chile), thus spanning more than half of the globe. The Austronesian, or, as it was formerly called Malayo-Polynesian language family, has its heartland in the Malay Archipelago (Indonesia, insular Malaysia and the Philippines, from a historic-linguistic point of view also including Taiwan), although some languages, most notably Cham, are native to continental Southeast Asia. As for the number of languages attributed to this family, close to 1,300 (and this may be underestimated), it is second only to the African Niger-Congo family. Official languages included in this family are Indonesian, Malaysian, Pilipino, Malagasi, and the national languages of the numerous new island states in the South Pacific (Samoan, Tongan, Fijian, Marshallese, Tuvalu, Kiribati etc.).

14. The Indo-Pacific languages

The languages of the islands of the South-western Pacific and North-eastern Indian Ocean area not belonging to the Austronesian family are grouped into the phylum of the Indo-Pacific languages. Kinship relationships between the different sub-phyla or families within this grouping are sometimes suspected, but have not (yet) been established.
The numerous languages of New Guinea and adjacent islands – more than 800 within Papua New Guinea, plus 250 in the Indonesian part of the island – are, as long as they are not Austronesian, classified into the *Trans-New Guinea* stock (550 languages); the *Sepik-Ramu* stock (100 languages); the *Torricelli* stock (50 languages); the East Papuan stock (35 languages); the *Geelvink Bay* stock (35 languages); the *West Papuan* stock (25 languages); and a number of minor stocks (7 *Left May* languages, 7 *Sko* languages, 6 *Kwomtari-Baibai*

languages, 3 *East Bird's Head* languages, 2 *Amto-Musan* and 2 *Lower Mamberamo* languages). Seven languages are considered 'isolates', i.e. not having kinship to any other language, while seven others remain unclassified..

15. Australian languages

As the aboriginal tribes of the Australian continent are believed to have dwelled in their southern homelands without disturbances from abroad for maybe as long as 40,000 years, the languages they make use of might be the oldest living languages on earth. Today still 250 of them are in use, only half the number estimated for the 18th century, and most of these are seriously threatened with extinction; as aboriginal societies crumble, their languages can nothing but dwindle with them. The Australian languages fall apart into 28 separate (sub) families, but all seem to be related. The northern one eighth of the continent shows the greatest diversity, all languages in the remaining area belonging to only one family called *Pama-Nyungan*.

16. Palaeo-Siberian languages

In the easternmost part of Siberia and the areas bordering it to the south, a number of small (in terms of numbers of speakers) unrelated language families are grouped together under the header 'Palaeo-Siberian'. The name of this grouping suggests antiquity, and indeed the languages belonging to these families are believed to have existed for long, and dominated a much more extensive area in the past than they do now. The families are (1) *Chukotko-Kamtchatkan* (also named Luorawetlan), of which Chukchi (12,000) and Koryak (8,000 speakers) are most prominent; (2) *Yukaghir*, at present still containing only the language with the same name (500 speakers); (3) *Yenisei Ostyak*, of which only Ket (1,000 speakers), spoken on the banks of the Central Yenisei river, is still convincingly alive; and (4) the language isolate *Gilyak* or *Nivkh*, spoken by 400 out of an ethnic population of 5,000 on the island of Sakhalin.

17. Amerindian languages

The languages spoken by the pre-Columbian societies of North and South America are currently classified in at least 50 separate language families. The largest of these are: (1) the *Oto-Manguean* family, consisting of a 170 languages, all but one Costa Rican outlier being spoken in Mexico; (2) the 75 *Arawakan* languages, ranging from Honduras, the Caribbean Islands and Surinam in the north to Argentina in the south; (3) the 70 *Tupi* languages spoken in Brazil, Peru, Bolivia, Paraguay and French Guyana; (4) the *Mayan* family, also represented with 70 languages, all spoken in the Yucatán Peninsula; (5) the *Uto-Aztecan* family, with a little over 60 languages extending from the western USA through Mexico into El Salvador; (6) the 47 *Quechuan* (Inca) languages spoken in the mountainous Andean area; (7) the *Na-Dene* family, with 42 languages represented from Alaska to the south-western United States; (8) the 33 *Algic* languages, consisting of the Algonquin subgroup, Wiyot and Yurok, al spoken in Canada and the USA; (9) the 32 *Macro-Ge* languages of Brazil and Bolivia; (10) the 29 *Panoan* languages of Brazil, Peru and Bolivia; (11) the 29 *Carib* languages, spoken from the area south of the Caribbean Sea into the Guyanas; (12) the 27 *Penutian* languages, spoken in the western USA and Canada; (13) the 27 *Hokan* languages of Mexico and the

south-western United States; (14) the 27 *Salishan* languages of Canada and the northwestern USA; (15) the 26 *Tucanoan* languages of Colombia, Ecuador and Brazil; (16) the 22 *Chibchan* languages spoken from Ecuador to the southern states of Central America; (17) the 17 *Siouan* languages of the Great Plains area of Canada and the USA; (18) the 16 Mexican *Mixe-Zoque* languages; (19) the 11 *Eskimo-Aleut* languages of the Arctic tundra's from Greenland to eastern Siberia; (20) the 11 *Mataco-Guaicuru* languages of Argentina, Paraguay, Bolivia and Brazil. Others include the *Iroquoian*, *Muskogean* and *Aymaran* families. Just like the language of the Australian aboriginals, the native American languages have spectacularly dropped in numbers as a result of Western colonization: more than 75% of the original number of languages may have disappeared.

18. Language isolates

Among the languages not (yet) accommodated in any one of the families that to date have been recognized, there are two important official languages: Japanese (125 million speakers) and Korean (75 million speakers). The Basque language, spoken in an area stretching over the western Pyrenean mountains in France and Spain, and having official status on a regional level in the latter country, is another example. A number of very prominent languages of the past, like Sumerian (the first known written language!) in Mesopotamia, Etruscan in Italy, and the language of the Mohendjo-Daro civilization in present-day Pakistan, and several pre-Indo-European languages of Europe (Iberian, Ligurian), are considered not to have (had) any relatives either
Of course there is something very dissatisfying about these so-called isolates. Assuming that mankind itself had a single origin – African, as we know believe - and different languages could only develop by estrangement caused by physical and hence social separation, the mere idea of languages isolates seems a bad excuse for our genealogic ignorance. Attempts are therefore sometimes made to assign isolates to established families anyway, and in the same attempt families themselves may be tentatively joined together into 'super-families'. Some suggest ties between Korean, Japanese and the Altaic family, that itself is by some thought to have a common background with the Uralic languages. In the 1980-s the American linguist Joseph Greenberg presented a new classification grouping together Indo-European, Altaic, Japanese, Korean and Eskimo-Aleut into one *Euro-Asiatic* family, forming part again of a super-family including also all Amerindian languages as well as quite a lot of isolates. That thus ceased to be isolates.[vi]

Back to toponymy

The quest for the real kinship ties between the languages is indeed an addictive intellectual pursuit, as it allows us to lift a corner of the veil of our own crepuscular past. To the practical toponymist, who pursues clarity about which rules should apply to which names, a difference is a difference: no matter whether it concerns dialects closely related to each other or languages belonging to unrelated families. It is the number of different languages/dialects one has to cope with that counts.

To get a hint what amount of linguistic knowledge is required in order to collect, record and standardize the geographical names within a single country, some statistics of the number of languages involved suffice. If we do not include the smallest independent states and territories, an average Asian country counts more than 60 native languages within its borders,

an African country about 50, an American over 40. Even in Europe, where national languages are known to have acquired a dominant position many centuries ago, the average country still counts seven languages. The Summer Institute of Linguistics counts more than 200 different native languages in as many as 11 countries: Papua New Guinea (822), Indonesia (729), Nigeria (513), India (397), Mexico (293), Cameroon (286), Australia (266), Brazil (232), the USA (227), the Democratic Republic of the Congo (219), and China (201). An additional nine countries (the Philippines, Sudan, Malaysia, Tanzania, Chad, Nepal, Myanmar, Vanuatu and Peru) count in between 100 and 200 languages, another 18 between 50 and 100. These numbers do not yet include languages classified by linguists as dialects.

[i] F.J. Ormeling Sr. – Terms used in geographical names standardization. In: T.R. Tichelaar (ed.), Proceedings of the Workshop on Toponymy held in Cipanas, Indonesia 16-18 October 1989. Cibinong, Bakosurtanal 1990.

[ii] B.F. Grimes (ed.) - Ethnologue: Languages of the World, Fourteenth Edition. Dallas, SIL (Summer Institute of Linguistics), 2002. Web-version: http://www.ethnologue.com/

[iii] M. Kolga, I. Tonurist et al. - The Red Book of the Peoples of the Russian Empire. Tallinn, 1993. Internet: http://www.eki.ee/books/redbook/

[iv] S.A. Wurm, S. Hattori (ed.) - Language Atlas of the Pacific Area. Canberra, Australian Academy of the Humanities, 1983.

[v] A. Room - Dictionary of Place-names in the British Isles. London, Bloomsbury 1988.

[vi] D. Crystal - The Cambridge Encyclopedia of Language. Cambridge, Cambridge University Press, 1987.