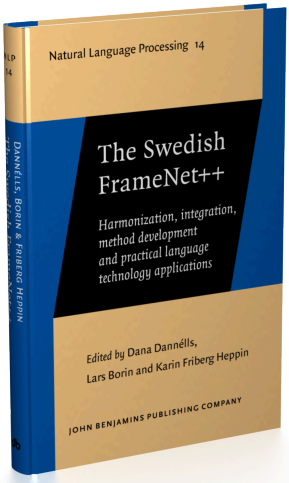
Natural Language Processing 14

CHAPTER 6

Swedish FrameNet++ and comparative linguistics*

Lars Borin | University of Gothenburg
Anju Saxena | Uppsala University
Shafqat Mumtaz Virk | University of Gothenburg
Bernard Comrie | University of CaliforniaSanta Barbara

https://doi.org/10.1075/nlp.14.06bor



Pages 139–166 of
The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications
Edited by Dana Dannélls, Lars Borin and Karin Friberg Heppin
[**Natural Language Processing**, 14] 2021. xiv, 333 pp.

CHAPTER 6

Swedish FrameNet++ and comparative linguistics[1]

Lars Borin[1], Anju Saxena[2], Shafqat Mumtaz Virk[1] and Bernard Comrie[3]
[1]University of Gothenburg / [2]Uppsala University / [3]University of California, Santa Barbara

In this chapter we describe a multilingual extension of Swedish FrameNet++, intended to address research questions of a broad comparative nature, in genealogical, areal and typological linguistics, focusing on the integration into Swedish FrameNet++ of so-called core vocabularies, used in several linguistic subfields in order to conduct massive comparative studies involving large numbers of languages. Specifically, we describe the inclusion of two such lexical databases covering several hundred South Asian languages, with the aim of investigating areal and genealogical connections among these languages.

There is no absolute judgment.
All judgments are comparisons of one thing with another.
Laming (1957: 9)

1. The multilingual aspects of Swedish FrameNet++

Swedish FrameNet++ has been a *contrastive* effort from its very beginnings – "contrastive" understood as involving comparison among a few – typically only two (van der Auwera 2012) – languages. The English Berkeley FrameNet (BFN) has served as a constant frame of reference throughout our work on the Swedish FrameNet (see Chapters 2 and 8 in this volume), and later the Multilingual FrameNet initiative (Torrent et al. 2018) as well as the work on linking SweFN++ to international wordnets (described in Chapter 5 in this volume), and the Swedish Constructicon initiative (Lyngfelt et al. 2018) have served as background for contrastive studies of frames and constructions. In this chapter, we describe a more profoundly

1. Parts of this chapter build on and elaborate content previously presented in Borin (2012) and Borin et al. (2013).

https://doi.org/10.1075/nlp.14.06bor

multilingual – *comparative* – aspect of SweFN++, prompted by some associated projects which have addressed research questions of a broad comparative nature, in genealogical, areal and typological linguistics. In particular, the notion of *core vocabulary* – or *basic vocabulary* – plays an important role in such investigations, thereby forming a natural link to SweFN++.

According to the standard language catalogues used by linguists – *Ethnologue* (Eberhard et al. 2021) and *Glottolog* (Hammarström et al. 2020) – there are about 7,000 languages in the world. *Comparative linguistics* is that branch of linguistics which investigates properties of and connections among these languages – genealogical, typological, geographical (areal), and universal – *based on empirical language data drawn from large and representative samples of the world's languages*.

As noted in the introduction to this volume (Chapter 1), the lexicon of a language is perhaps its most salient characteristic and the most obvious expression of the connection that language bears to the world. The lexicon is where phonology, grammar, semantics, and pragmatics come together in language, and in some sense, language knowledge *is* vocabulary knowledge. Importantly, the lexicon also reflects the genetic affiliations of a language and its contact history, and it can be used to elucidate language change, both in meaning and in grammar.

This has led students of comparative linguistics to compile lexical databases containing comparable lexical items and their associated linguistic information from large numbers of languages. Since the work in SweFN++ has been conducted in close proximity to such research projects, and since SweFN++ itself already as originally conceived was to include a multilingual part, it became natural to partly merge these initiatives in order to avoid duplication of effort and to ensure extensibility of the resulting resources. Furthermore, the broad comparative perspective introduces a healthy and stimulating outsider point of view also on monolingual lexical description.

Below we describe the two multilingual lexical databases developed in comparative linguistic projects associated with SweFN++, and some technical and conceptual issues connected to their inclusion in SweFN++. However, first we will need to discuss the notion of *core vocabulary* which becomes central in this context.

2. Core vocabularies for comparative linguistic studies

2.1 Basic vocabularies in linguistics

The notion *core vocabulary* plays a significant role in several linguistic subdisciplines, but with different meanings and based on different theoretical and methodological premises. Taking a step back and trying to abstract away from irrelevant detail, we observe that what is "basic" or "core" about core vocabularies is broadly speaking construed in three ways in the linguistic literature.

2.1.1 *Semantic simplicity*

In *lexicology* and *lexicography*, the core vocabulary is equivalent to a *defining vocabulary*: a set of words using which all definitions in a dictionary must be expressed, directly or indirectly, and consequently, these words themselves will not be defined – only described – in the dictionary. This is in principle a language-specific notion; different languages could have different core vocabularies. Good examples are provided by English learner dictionaries, e.g., the approximately 2000-item *Longman Controlled Vocabulary* used in the definitions in the *Longman Dictionary of Contemporary English* (Xu 2012). The SweFN++ pivot resource Saldo is also organized along these lines (see below and Chapter 3 in this volume).

In *semantics*, as a language-independent extension of the foregoing, the core vocabulary is a set of senses – universal lexical-semantic primitives – in terms of which all vocabulary items in all languages can be expressed (Wierzbicka 1996; Goddard 2008). On some construals, however, these primitives need not actually correspond to lexical items in any language.

In recent years, the field of *language technology* has been added to the above. Here, core vocabularies enter the stage in the form of ("upper") *ontologies*, i.e., formally organized hierarchical concept systems (Huang et al. 2010). These systems are not explicitly made up of lexical or even linguistic items, but in practice they appear as a kind of lexical structures, as noted by Wilks (2009: 4):

> [I]tems in ontologies and taxonomies are and remain words in natural languages – the very ones they seem to be, in fact […] Those who continue to maintain that 'universal words' are not the English words they look most like, must at least tell us which of the senses closest to the 'universal word' they intend it to bear under formalisation.

2.1.2 *Early acquisition/commonness/representativeness/frequency*

In *applied linguistics*, corresponding to the notion that vocabulary growth in language learners is not random or spurious, the core vocabulary is the vocabulary appearing first in L1 acquisition, and the vocabulary most useful or central in L2 learning, the most basic set of words that a language learner will need to master in a foreign language in order to fulfill some minimal requirement of competence in the language. In an activity associated with SweFN++, the KELLY project,[2] the aim was to develop vocabularies corresponding to the six language learner proficiency levels of the *Council of Europe's Framework of Reference* (CEFR),[3] amounting to 1,000–2,000 words per level. The vocabularies were developed for nine languages –

2. https://spraakbanken.gu.se/en/projects/kelly

3. The levels are designated (from beginner to advanced) A1, A2, B1, B2, C1, and C2.

Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian and Swedish – and translated among all language pairs by professional translators (Kilgarriff et al. 2014). Here, the notion of coreness corresponds to adjusted frequency in a very large corpus collected from the WWW using the Web-as-Corpus methodology (Baroni & Bernardini 2006). The lists for the lowest CEFR level containing approximately 1,500 vocabulary items are arguably candidates for core vocabularies for these languages.[4]

In *corpus linguistics*, a kind of core vocabulary are the most frequent words evenly dispersed over a broad range of text types (Forsbom 2006).

Ogden's (1930) *Basic English* and other forms of simplified language, such as Voice of America's (2009) *Special English* or the many forms of *controlled natural language* developed for special LT purposes (Kuhn 2014), often come with a vocabulary made up of the "most often used words" (Kuhn 2016: 102).

2.1.3 *Resistance to replacement*

In *historical-comparative linguistics*, core vocabularies are brought to bear on the question of genetic classification of languages. Since the time depths involved are more often than not counted in millennia, core vocabularies should be composed of words known to be resistant to replacement even over long time periods. This was the explicit motivation for the well-known *Swadesh lists*, first proposed in the late 1940s by Morris Swadesh (1948, 1950, 1952, 1955) – (short) lists of supposedly universal concepts together with their lexical realizations in many languages, which are used in an endeavor referred to as *lexicostatistics*, the primary purpose of which is to investigate genetic relationships among related languages.

Lexical items can be replaced by language-internal elements, but also through borrowing, which is studied in areal (or contact) linguistics. Thus, one kind of core vocabulary can be defined more narrowly as being made up of lexical items resistant to borrowing, e.g. the Leipzig-Jakarta list described by Haspelmath & Tadmor (2009).

2.1.4 *Related kinds of vocabularies*

Lexical data from many languages collectively are the object of study of the branch of linguistics known as *lexical typology*, where the limits of the lexical diversity of human languages are studied (Koptjevskaja-Tamm 2012; see also Evans 2011). For comparability, a set of meanings with universal or near-universal lexicalization (e.g., Goddard 2001, 2012) are posited, for investigating questions such as the universality of lexical expression, frequently attested types of semantic change, borrowability, etc.

---

4. Although in KELLY the most frequent items are actually excluded as belonging to an even more basic, "sub-CEFR", vocabulary level.

Further, the literature on experimental psychology is also full of descriptions of related kinds of word lists, usually referred to as "norms", ordered by i.a. frequency, but other commonly used parameters are also age of acquisition, concreteness and emotional or sentiment value (often called *polarity* in this context). A reason to discuss such norms in this connection is that, in order to work as intended, the lexical items in them must be well-known to experimental subjects, i.e., they must be common words, one of the criteria for coreness. There is also reason to believe that the various criteria for core vocabulary membership discussed in the linguistic literature are intercorrelated to an unknown extent, and that at least some of the properties defining psycholinguistic norms also belong here.

2.2 The composition and size of core vocabularies

2.2.1 *The "words" of core vocabularies*

In the literature on core vocabularies, the lexical items by which they are composed are – again speaking broadly and referring to actual practice – of three kinds, *concepts, word senses* and *lexemes*, where the latter two are language-specific while concepts generally are considered to be independent of language.

The kinds of core vocabularies in focus here – those conceived in the context of large-scale comparative linguistics – are always described as being composed of concepts. The interesting question – which is mostly consigned to silence in the literature – is how concepts in such core vocabularies and lexical units in dictionaries of actual languages are ontologically, as it were, interconnected.

It is far from clear exactly what a "concept" is in this literature. Are the available concepts the union of those which find lexical expression in some – at least one – language out of the world's approximately 7,000 languages? Or is the inventory of concepts independent of language, so that there will be concepts that never receive linguistic expression in any language? Logically, this independence must be one-way, however; it seems that everything that gets lexical expression in at least some language, will also necessarily be a concept.[5] The literature on concepts – in linguistics, philosophy and psychology – is actually too vague to be of much use to us in our ontological quandary. It is sometimes proposed in the literature that concepts have compositional structure, a bit like many linguistic expressions, including

5. As we will also see below, we don't need to look long in the world's languages in order to find lexical units with "funny" meanings (i.e., strange from the point of view of more well-described languages). Evans & Levinson (2009: 435) provide the example of the Mundari (unr; an Austroasiatic language spoken in South Asia) ideophone *rawa-dawa*, which they gloss as 'the sensation of suddenly realizing you can do something reprehensible, and no-one is there to witness it', which consequently should count as a concept. This kind of example can be repeated essentially *ad infinitum*.

the words typically used as labels for concepts, but it seems that this idea has not been systematically pursued independently of language, undoubtedly because it is very difficult to discuss concepts independently of the words used to express them.

In the introduction to the Princeton WordNet volume (Fellbaum 1998b) we read that "[t]he majority of lexicalized concepts are shared among languages" (Fellbaum 1998a: 8). The findings reported in recent work in linguistic typology and language universals clearly run counter to this claim, which shows the need for more research into this matter:

> languages do differ almost without limit as to which meanings they choose to lexicalize (von Fintel & Matthewson 2008: 151)
>
> languages differ enormously in the concepts that they provide ready-coded in grammar and lexicon [… and] many languages make semantic distinctions that we certainly would never think of making (Evans & Levinson 2009: 435)
>
> There is no sense of "broad" under which "the grammars and lexicons of all languages are broadly similar." (Levinson 2003: 28)
>
> From even a small sample of languages it is clear that many impressionistically "basic" items of English vocabulary (such as *go, water* and *eat*) lack exact equivalents in other languages. (Goddard 2001: 57)

2.2.2 *Selecting core vocabulary concepts*

The lexicon of any language is vast, while a common characteristic of many core vocabularies is that they are small, especially those used in comparative studies. The Swadesh list has actually shrunk over time. From an original size of about 200 items (Swadesh 1952), it was soon pared down to the widely used 100-item version by the elimination of items judged to be not universal enough or not fully arbitrary (e.g., onomatopoetic or sound-symbolic), etc. (Swadesh 1955). The most recent Swadesh-style lists are much shorter. Thus, the ASJP (Automated Similarity Judgement Program) list (Holman et al. 2008), holds only 28–40 items, and the list presented by Dolgopolsky (1986) contains 26 items.

How do we pick out the most "core" lexical items for inclusion in a core vocabulary? In particular, can we select a set of core concepts for the large-scale comparative studies that are in focus in this chapter? In the literature we find descriptions of experiments conducted in order to estimate the usefulness of individual core vocabulary items in some candidate list given in advance (e.g. Holman et al. 2008). In information retrieval terms, what is discussed and investigated in these cases is *precision*: How many of the proposed list items ought to be in the list for it to be usable for the intended purpose? The complementary measure of *recall* is not systematically scrutinized: how many items are missing from the list which ought to be in it (and which are those items, and more importantly: how do we recognize them)?

Here, the initial basis for choosing which items to include becomes crucial. Ideally, any list should be assembled using the full lexicons of all included languages (or of a large representative sample of languages in the general case) as the sampling frame. As far as we can see, none of the popular extant basic vocabularies lives up to this ideal, with the possible exception of the Natural Semantic Metalanguage (NSM; Goddard 2012). For instance, both the ASJP list (Holman et al. 2008) and the Leipzig-Jakarta list of basic vocabulary resulting from the LWT project (Loanword Typology; Tadmor 2009) are based on much smaller, judgment-based sampling frames. The ASJP list is made up of the most stable items from the 100-item Swadesh list, and the Leipzig-Jakarta list is made up of the 100 items most resistant to borrowing from the slightly extended IDS (Intercontinental Dictionary Series) list used in the LWT project. In both cases it is probably fair to say that the original selection of items was made "by a combination of intuition and experience following certain guidelines" which characterized Swadesh's work (Oswalt 1971: 422). For the predecessor of IDS, Buck (1929, 1949) does not explicitly state the criteria for which items should be included, beyond the goal to "work out a tentative and skeleton dictionary covering a limited number, perhaps a thousand, of representative groups of synonyms in the principal I[ndo-]E[uropean] languages" (Buck 1929: 216). Thus, in both cases, we have no hard empirical evidence that the ASJP list and Leipzig-Jakarta list comprise the optimal selection of items in their respective categories. There could in principle be a better 100-item Swadesh list or a better 1460-item LWT/IDS list. The overlap between the 100-item Swadesh list and the (100-item) Leipzig-Jakarta list is only 62% (Tadmor 2009: 73) and only about a third of the 42 NSM items listed by Goddard (2001) are present in either of these two lists (Borin 2012). This indicates that there is still much to be found out about basic vocabularies.

2.2.3 *Selecting core-vocabulary lexemes*

In the absence of an established formalism for expressing concepts unambiguously, natural-language words are used – typically English words and phrases – to indicate the concepts in core vocabularies. Tentatively we may still compare core vocabularies simply using the English glosses and assume that they in the normal case – especially for the small vocabularies that we will be concerned with here[6] – reflect the same or at least comparable senses, analogously to how color words have been suggested to share central/focal meanings across languages even if their boundaries vary greatly cross-linguistically (Berlin & Kay 1969; Kay & McDaniel 1978). However, in the case of color terms, there are demonstrable physical and

6. Since several of the vocabularies ultimately come out of the same research tradition, we may tentatively assume commensurability at least for the items in those lists.

physiological features that can be adduced to explain this. This becomes a trickier proposition in the general case, because of the difficulties inherent in defining a language-neutral set of concepts for those many cases where there is no physical or concrete object to use as basis for establishing the prototype or central meaning.

However, we cannot ignore the circumstance that this introduces a potential colexification issue. Many of the words used to express proposed core vocabulary concepts are polysemous, and the linguist or language consultant supplying the concrete vocabulary items corresponding to the concepts in a particular language will need to rely on their intuition in order to pick the right alternative. The remark by Yorick Wilks quoted earlier puts the finger squarely on a practical-methodological difficulty which arises in connection with the comparison of core vocabularies: How do we determine that two vocabulary items are equivalent, in one language and – in particular – across languages? In other words: Do we know how to compare vocabulary items?

A complementary issue to the preceding is synonymy. Even if a language consultant picks out the intended concept, there may still be more than one way of expressing it, and there is some experimental evidence that native speakers will collectively produce more than one alternative for many items (Slaska 2005). This has practical consequences for how such proposed core vocabularies can be used in comparative linguistic investigations.

2.2.4 *Comparing core vocabularies*

An interesting question in the context of core vocabulary compilation is how correlated the various defining properties discussed above are. In the literature we find both more speculative and more empirical attempts to resolve this question.

To make things more concrete, in Table 1 we show six different core vocabulary lists (each list ordered alphabetically):

1. The Automated Similarity Judgement Program (ASJP)[7] 40-item vocabulary for genetic and areal linguistics research (Holman et al. 2008) (referred to as A40 below)
2. The first 40 items of the Leipzig-Jakarta (LJ) vocabulary from the loanword typology project[8] (Haspelmath & Tadmor 2009) (L40)
3. Goddard's 42 universal lexical items (Goddard 2001) (G42)

7. https://asjp.clld.org/

8. https://wold.clld.org/; the LJ list has been somewhat edited for the purposes of this comparison, so that, e.g., "1sg pronoun" has been replaced by "I", "arm/hand" by "hand", "who?" by "who", "child (kin term)" by "child", etc.

Table 1. Six different core vocabularies compared

A40 (ASJP)	L40 (LWT)	G42 (Goddard)	S41 (Saldo)	B42 (BV pool)	K-8 (40 items) (KELLY)
blood	arm	above	about	about	bomb
bone	big	after	all	a(n)	bread
breasts	bitter	all	be	and	bridge
come	blood	a long time	before	as	channel
die	bone	bad	but	be	climate
dog	breast	because	by	be able/can	coffee
drink	come	before	color	become	dog
ear	do	below	do/make	big/large	eye
eye	ear	big	exactly	but	father
fire	far	die	for	come	forest
fish	fire	do	good/well	do/make	future
full	fish	good	happen	exist	government
hand	flesh	happen	have	for	hospital
hear	fly	hear	how	from	kitchen
horn	foot	here	if	get	knee
I	go	I	in (prep.)	have	level
knee	hair	if	in/at front	he	library
leaf	he/she	inside	know	I	logic
liver	hit/beat	know	light (noun)	if	marriage
louse	horn	like	move	in	member
mountain	house	live	much	it	minister
name	I	maybe	must	not	music
new	louse	much/many	name	of(f)	office
night	mouth	not	nature	on	pocket
nose	name	now	on	one/they [3.IMPERS]	poet
one	neck	one	only	or	prison
path	night	people	other	other	problem
person	nose	say	quickly	REFL	queen
see	one	see	say	REFL.POSS	revolution
skin	rain	small	sound (noun)	shall/will	sand
star	root	someone	straight	she	source
stone	say	(some)thing	think	so	sun
sun	stone	there is	to (prep.)	that (subjunction)	tea
tongue	this	the same	want	the	theory
tooth	tongue	think	warm/hot	they	third
tree	tooth	this	what	this	trade
two	water	two	when	to (prep.)	tragedy
water	who	want	where	to (prep. w. inf.)	university
we	wing	very	who	we	water
you	you	when/time	with	when	week
		where/place	yes	which/that	
		you		with	

4. Saldo's 41 top-level word senses translated into English; see Chapter 3 in this volume for the Swedish items (S41)
5. The first 42 items in Forsbom's (2006) *base vocabulary pool* (excluding punctuation marks), translated into English (BV pool). The BV pool has been computed on the basis of a one-million word balanced corpus of Modern Swedish (Gustafson-Capková & Hartmann 2006) (B42)
6. All items common to at least eight of the nine KELLY languages (40 items) (K-8)

The vocabulary that sticks out in the table is K-8. As mentioned above, even the lowest level of KELLY vocabularies excludes the most high-frequent items as being too basic even for CEFR level A1. This is clearly seen if we compare the other vocabulary lists with the full English KELLY A1-level list (1250 items). These comparisons yield from one to three items in common. If we instead make the same comparisons with the 800-item Basic English list (Ogden 1930) or the first 2000 items in the SUBTLEXus word-form frequency list derived from a corpus of subtitles in films (Brysbaert & New 2009),[9] we get on the order of 70–80% correspondences (percentages in terms of the short 40–100 item lists), which is more like what we would expect of a basic vocabulary list. Still, it is noteworthy that only 80% of the A40 list and 76% of the 100-item ASJP list are found in these long basic vocabulary lists for English, i.e., a fifth to a quarter of the ASJP items are missing. ASJP sense labels are English words, after all, so somehow one would expect a figure much closer to 100%.

The K-8 list is thus not comparable to the ASJP or LJ lists, but since it reflects a set of common words that have emerged from a combination of corpus processing and manual translation among all 72 language pairs, a comparison of K-8 with G42 should arguably be meaningful.[10]

Further, A40, L40 and G42 have three items in common – *I, one* and *you* – and the longer 100-item ASJP and Leipzig-Jakarta lists share 12 items with G42. G42 has no common items with the K-8 list, and shares only three items with the longest 271-item list of items common to seven KELLY languages: *big, time* and *two*.

2.2.5 *Conclusions*

It is perhaps not surprising that there should be so little overlap among different kinds of core vocabularies, since they aim at capturing different aspects of coreness:

9. https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus

10. Here the thorny issue of translation equivalence rears its head. The KELLY translators were instructed to provide *one* translation in the normal case. How this methodological decision has influenced this investigation is a very interesting question which we will have to leave for future research.

- **Item stability:**
  - The ASJP vocabulary consists of maximally stable form–meaning pairs; it can be seen as a refinement of traditional Swadesh lists, based on a broad range of empirical data.
  - The Leipzig-Jakarta list reflects resistance to borrowing.
- **Sense inventory:**
  - Goddard's list contains senses that are highly likely to receive lexical expression in all languages.[11]
  - Saldo's top-level word senses are those that are "most central" according to the Saldo criteria.
  - The BV pool and basic learner vocabularies contain high-frequency – much-used and consequently highly useful – lexical items.

Still, we would expect the ASJP list to be a subset of the LJ list, instead of showing a not too large overlap with it, which is what we actually find. This is because resistance to borrowing is a special case of resistance to vocabulary item replacement. Another common form of replacement is the substitution of a native vocabulary item by another native word.

There is no logical need for universal word meanings to be highly frequent. Highly technical vocabulary could be expected to behave in a way that could make it universal in Goddard's sense – because such vocabulary items would mean the same wherever they occurred. In this sense Goddard's criteria and those used in compiling learner vocabulary lists are orthogonal to the criterion of replacement. However, there may be good pragmatic reasons which in practice single out more or less the same sets of items. Highly technical vocabulary – regardless of whether it belongs in the realm of rainforest botany or solid-state electronics – will be confined to a small fraction of the world's languages in each individual case. Hence, Goddard's universal word meanings will by pragmatic necessity belong to everyday language.

High-frequency senses may or may not undergo the linguistic equivalent of an extreme makeover. From experience we know that, e.g., sentence adverbs and indefinite pronouns – arguably central and in part universal vocabulary items – often are non-cognate even in closely related languages.

On the other hand, it is often mentioned in works on historical linguistics that high-frequency items tend to preserve older inflectional patterns as irregularities (seen from the point of view of the present-day inflectional system), which can consequently be used in internal reconstruction for inferring the older system. Intuitively, the older inflectional patterns should be accompanied by the corresponding older lexical items, i.e., this would lead to the conclusion that high-frequency

11. As we have seen above, this may be true only in an approximate sense.

vocabulary should be more stable than the above-mentioned comparisons with the Basic English and SUBTLEXus lists show.

Why the ASJP and LJ lists do not show the expected inclusion relationship and why high-frequency central vocabulary is less stable than expected are two mysterious aspects of core vocabularies which will need further investigation.

3. Two lexical databases for investigation of South Asian linguistic diversity and unity

3.1 Linguistic diversity in South Asia

South Asia (also India[n subcontinent]) with its rich and diverse language ecology and a long history of intensive language contact provides abundant empirical data for studies of linguistic genealogy, linguistic typology, and language contact.

This region (normally understood in linguistic works as comprising the seven countries Bangladesh, Bhutan, India, the Maldives, Nepal, Pakistan, and Sri Lanka, as well as adjacent areas in neighboring countries, since language boundaries do not always coincide with national borders) is the home of hundreds of languages spoken by almost two billion people – more than a quarter of the world's population. Most of the some 700 living languages of South Asia (Eberhard et al. 2021) are from four major language families: Indo-European > Indo-Aryan and Nuristani, Dravidian, Austroasiatic > Munda, Khasian and Nicobaric, and Sino-Tibetan (also called Tibeto-Burman and Trans-Himalayan); see Figure 1.[12] In addition there are some language isolates and small families (Georg 2017) and several creoles and pidgins.

At least since the publication of Emeneau (1956), South Asia has been considered a prototypical *linguistic area*, a region where, due to close contact and widespread multilingualism, languages have influenced one another to the extent that both related and unrelated languages are more similar on many linguistic levels than we would expect.

However, systematic investigations of this claim have been scarce (e.g., Masica 1976), mostly relying on data from a few major Indo-Aryan and Dravidian languages, but more rarely the other language families of South Asia (Ebert 2006). The picture of which areal phenomena are characteristic of South Asia, as well as of their geographical extent, is actually far from clear (Thomason 2000).

---

12. Source: Wikimedia Commons https://commons.wikimedia.org/wiki/File:South_Asian_Language_Families.jpg, license: CC BY-SA 3.0 Unported, consulted on 2019-10-12. Note however that the map incorrectly indicates as "unclassified/language isolate" the two small language families of the Andaman Islands, Great Andamanese and Ongan.

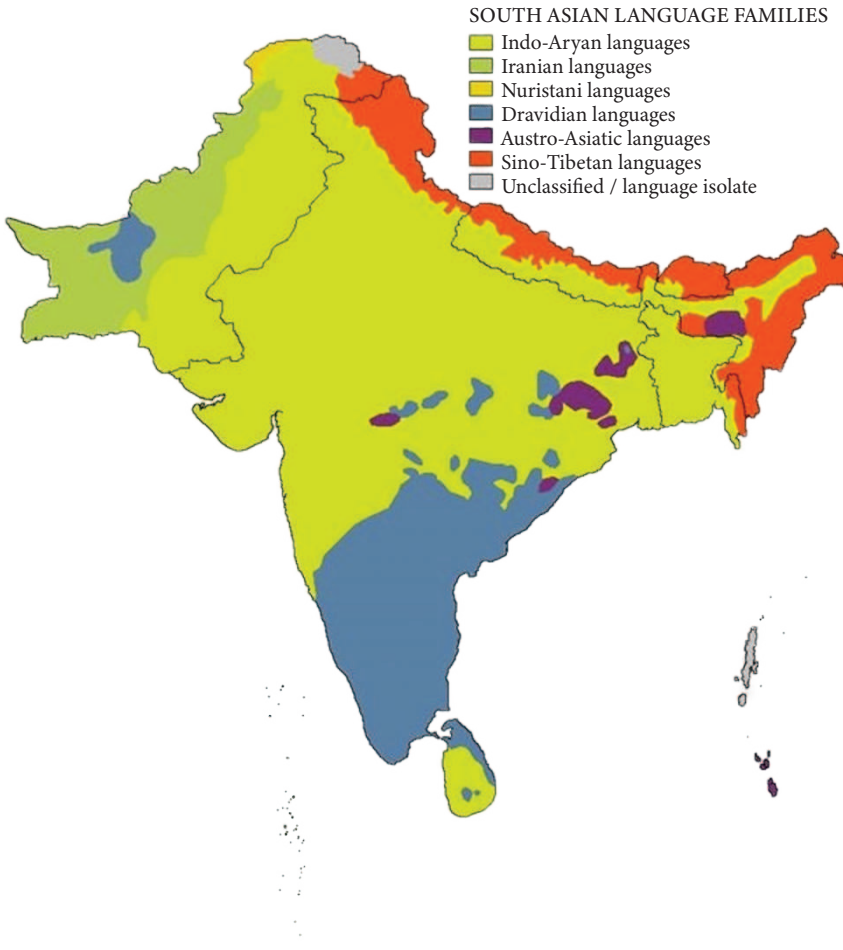


Figure 1. South Asian language families (source: Wikimedia Commons)

3.2 Grierson's *comparative vocabulary* in Swedish FrameNet++

3.2.1 *The* Linguistic survey of India

The linguistic richness and diversity of South Asia was documented by the British government in a large-scale survey conducted in the late nineteenth and the early twentieth century under the supervision of Sir George Abraham Grierson and Sten Konow. The survey resulted in a detailed report comprising 19 volumes of around 9,500 pages in total, entitled *Linguistic survey of India* (LSI; Grierson 1903–1927). The survey covered 723 linguistic varieties representing the major language families of the region and some unclassified languages, of almost the whole of

nineteenth-century British-controlled India (modern Pakistan, India, Bangladesh, and parts of Burma). For each major variety it provides a *grammar sketch* (including a description of the sound system); a *core word list*; and *text specimens* (including a morpheme-glossed translation of the *Parable of the prodigal son*).

The LSI grammar sketches provide basic grammatical information about the languages in a fairly standardized format. The focus is on the sound system and the morphology (nominal number and case inflection, verbal tense, aspect, and argument indexing inflection, etc.), but there is also some syntactic information to be found in them. The language data for the LSI grammar sketches were collected around the year 1900, hence obviously reflecting the state of these languages of about a century ago. However, we know that many grammatical characteristics of a language are quite resistant to change (Nichols 2003), much more so than vocabulary. Despite its age, LSI still remains the most complete single source on South Asian languages.

For this reason, we initiated a project some years back, entitled *South Asia as a linguistic area?*, with the aims of digitizing the full text of the LSI and developing LT tools for extracting linguistic features from the text in order to populate a typological database covering several hundred South Asian languages, which would then be used to investigate areal and genealogical connections among these languages.

As a concrete outcome of the project, most of the LSI is available for full-text search through Korp, the corpus tool maintained by Språkbanken Text at the University of Gothenburg (Borin, Forsberg & Roxendal 2012).[13] In addition to providing a user-friendly search and browsing interface to the full text of the LSI, it has acquired a particular significance to the SweFN++ endeavor. One of the LT methods devised for extracting linguistic features from the LSI grammar sketches is *frame-semantic parsing* using a domain-specific linguistic framenet (for English): *LingFN* (Malm et al. 2018). The corpus search engine has turned out to be indispensable to this work, allowing us to test hypotheses about linguistic frames quickly. See Figure 2, showing the result of a search for occurrences of the verb *place* in the LSI.[14]

---

13. A scanned version of LSI is available on the University of Chicago's *Digital South Asia Library* (DSAL) website – http://dsal.uchicago.edu/books/lsi/ – although the page images displayed there are neither searchable nor digitally processable, effectively making this version equivalent to the printed LSI wrt accessing its contents, although of course universally accessible to anybody with an internet connection. The Korp search interface provides a link to the corresponding DSAL page image for each hit (see Figure 2).

14. The text of the LSI has been linguistically annotated with lemma, part of speech and dependency syntax using the Stanford CoreNLP toolkit (Manning et al. 2014), so that the search shown in Figure 2 was done by specifying lemma ("place") and the first part of the part of speech tag ("VB") in Korp's extended search interface.

Figure 2. Korp view showing search results for "place + VB.*" in LSI

The LSI grammar sketches contain large amounts of tabular material, e.g., inflection tables, personal pronoun systems, etc., which are not suitable for displaying in a corpus KWIC (key word in context) view. Instead, these are imported and stored in another of Språkbanken's infrastructure components, itself an outcome of the SweFN++ initiative, Karp (Borin, Forsberg, Olsson, et al. 2012; Ahlberg et al. 2016; see also Chapters 1 and 2). Links are provided from the Korp KWIC metadata box to tables and specimens in Karp, but these can also be accessed directly through the Karp search interface. See Figure 3, illustrating a query aiming at finding out some linguistic features of the personal pronominal systems of the LSI languages.

3.2.2 *The LSI comparative vocabulary*

The focus of this chapter is Grierson's *Comparative vocabulary*, a separate LSI volume (Grierson 1903–1927: V1P2) collecting the core vocabularies which accompany the language descriptions, where we find wordlists for 240 South Asian language varieties, and also for some languages outside South Asia. Each list holds a total of 168 entries. Most of the entries in the comparative vocabulary render concepts which cover a broad spectrum consisting of body parts, domestic animals, personal pronouns, numerals, and astronomical objects. There is some overlap with other concept lists used in language classification: For instance, 38 of the concepts are also found in the shorter (100-item) Swadesh list. Thus, the LSI comparative vocabulary clearly has one part that can be used in investigating genetic connections among the languages, but also another part – at least half of the entries – which could be used to find areal influences.

The LSI comparative vocabulary is similar in spirit and extent to a Swadesh list, but of course predating the latter by half a century. In fact, Grierson had adopted

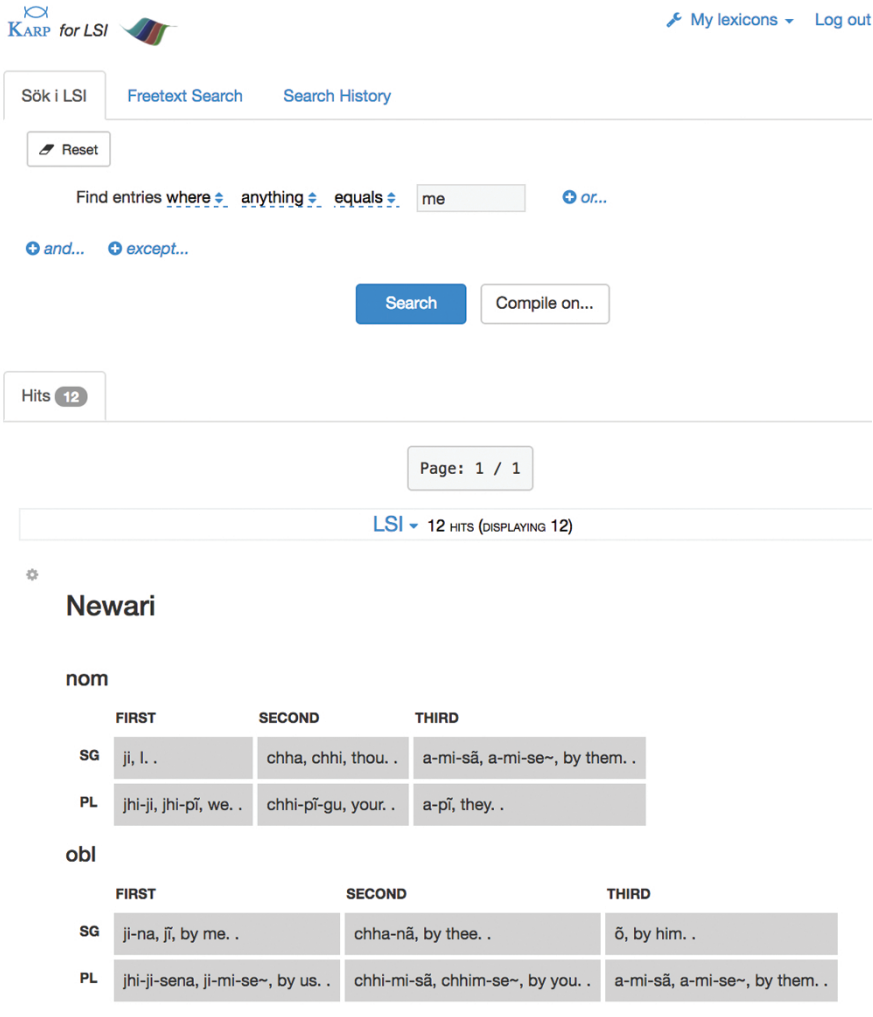


Figure 3. Karp view showing LSI tables of personal pronoun paradigms

the list from an even earlier source, viz. Campbell (1866b), whose *List of words and phrases to be noted and used as test words for the discovery of the radical affinities of languages, and for easy comparison* (Campbell 1866a) formed the basis for the LSI comparative vocabulary (Grierson 1903–1927: V1P1:17). Campbell motivates his selection of items in a way which could equally well have been formulated by Swadesh almost a century later:

> There are certain words which may almost be taken as unfailing tests in classifying language ; for instance, the first few numerals, the names for the commonest parts of the human body – as hand, foot, nose, eyes, mouth, head, &c. – the names of the commonest family relations – father, mother, brother, sister – sun and moon, fire and water – the personal pronouns, and one or two others. […] I shall also make a smaller list of English words, a translation of which I would recommend to be sent with each account of a tribe or race, speaking a language in any degree peculiar.
> (Campbell 1866b: 8f)

A subtle but important difference compared to typical Swadesh-style core vocabularies: Both Campbell's vocabulary and the LSI comparative vocabulary provide *morphosyntactic words* (Haspelmath 2011) instead of (or in addition to) concepts. Presumably this is the intention of entry glosses such as 'of a father', 'mares', etc. Both vocabularies also provide some phrases and propositions (e.g., 'good man' ~ 'good woman' ~ 'good men' ~ 'good women', and 'I, thou, etc. go' ~ 'I, thou, etc. went'), thereby representing a cross between a Swadesh list and the questionnaire format often used in modern typological data collection. This makes the LSI comparative vocabulary useful for comparative studies of some grammatical features, in addition to studies of more purely lexical phenomena. In a preliminary study, some grammatical features have been semiautomatically extracted from the comparative vocabulary. See Figure 4 where the feature *order of cardinal numeral and noun* is shown on a map of South Asia, illustrating that this feature has both genetic – Sino-Tibetan tends to have the order noun–numeral – and areal features – even some Sino-Tibetan languages have the order numeral–noun in the western part of the region.[15]

3.3 The *Intercontinental Dictionary Series* as a comparative linguistic research tool

3.3.1 *A lexical basis for investigating the linguistic landscape of the indian Himalayas*

In two other associated projects we have investigated the linguistic landscape of the Indian Himalayas, and adjacent areas in Pakistan and Nepal, with a special focus on the genealogy of and prehistorical contacts among the local Sino-Tibetan and Indo-Aryan language varieties.

---

15. Figure 4 is generated by a visualization tool developed in our project (see https://spraakbanken.gu.se/en/projects/digital-lsi/tools-and-resources, under "Static maps").

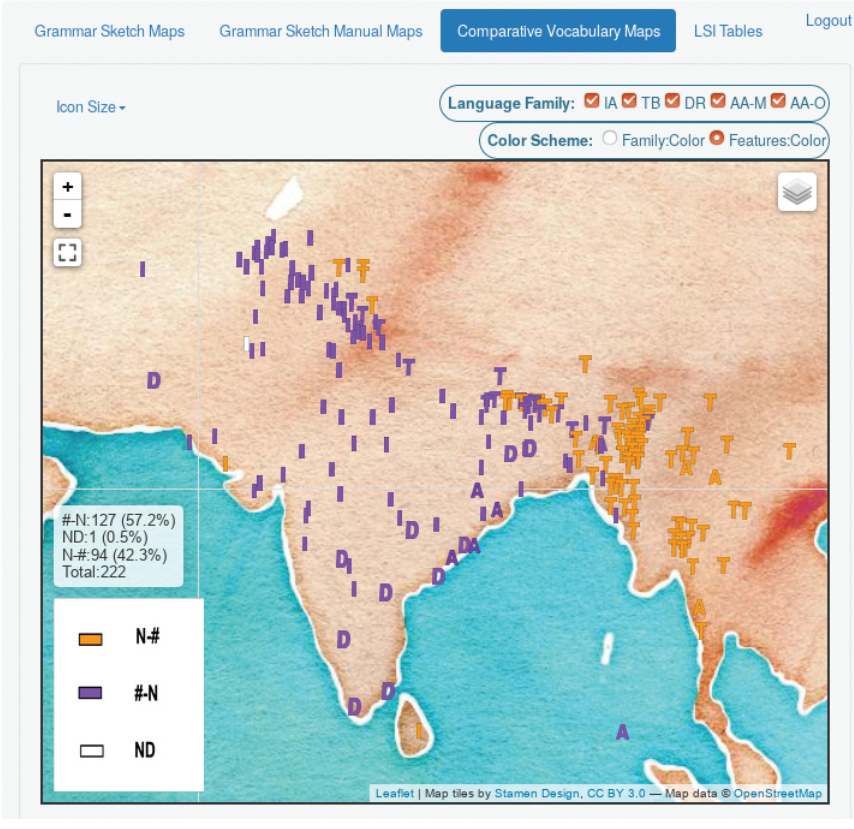


Figure 4. The feature *order of numeral and noun* as extracted from the LSI comparative vocabulary (Legend: A: Austroasiatic; D: Dravidian; I: Indo-Aryan; T: Sino-Tibetan [Tibeto-Burman])

An important goal of the *Digital areal linguistics* project was to create a database of comparable lexical items in a number of representative South Asian languages, with a focus on the Himalayan region in India and to use this database for investigating the Himalayas as a linguistic area. Long-standing contact between different language families, as well as among different subbranches of the same language family (for example, within the Sino-Tibetan language family) in the Himalayan region has resulted in intense lexical and grammatical borrowing. For languages where we do not have historical data, it is hard, or sometimes impossible, to distinguish similarities between languages due to common heritage from those due to contact. Based on lexical, grammatical and cultural data collected in the projects and the information available in secondary sources, we have been able to conduct a systematic comparative study of the lexical domain (phonology, morphology and

lexical semantics), as well as selected morpho-syntactic constructions – which lie at the root of areal linguistics (Thomason & Kaufman 1988; Heine & Kuteva 2005; Matras & Sakel 2007) – with a goal to investigate the areal hypothesis.

As the basis for this lexical database we selected the *Intercontinental Dictionary Series* (IDS), an international collaboration for establishing a database where lexical material from a broad range of languages is organized in such a way as to provide a solid quantitative base for a scientific approach to language analysis and comparisons.[16] Historical studies, comparative, and theoretical linguistic research can be based on this documentation.

IDS is a long term cooperative project that involves linguists all over the world. It is a pioneering effort that contributes to preserving information on the little-known and non-prestigious languages of the world, many of which are becoming extinct. The project brings together data on the languages of the world, in a way that gives equal importance to all languages.

The originator of the IDS was the late Mary Ritchie Key at the University of California, Irvine. The idea for a work such as the IDS came to her in 1975 while studying the semantic grouping in the cognate sets established in comparative studies. This was followed by pilot projects using comparative data of recognized language families. In 1984, an award from the University of California, Irvine Faculty Research Committee to launch the IDS set the series on its way. Bernard Comrie (University of California, Santa Barbara) is the current project leader of IDS and the general editor of the series.

The organization of IDS is modelled on *A dictionary of selected synonyms in the principal Indo-European languages* (Buck 1949). This 1500-page dictionary is organized in a thesaurus-like topical outline of 22 chapters. The outline has been adapted for the IDS, with the numbering system generally maintained. Buck's dictionary contains approximately 1,200 potential entries. The IDS adaptation contains 1,310 entries. The entries are identified using English words as labels, but they are intended to represent (lexicalized) concepts falling roughly into the following categories:

1. universal concepts finding expression in most human languages ('arm', 'speak', 'dry', etc.; but cf. Section 2.2.1 above and Goddard 2001);
2. concepts related to certain geographical or environmental phenomena: 'earthquake', 'tide', 'parrot', etc.;
3. cultural concepts: 'mead', 'tattoo', 'cobbler', etc.

---

16. In this section we make use of some introductory text on the IDS prepared by the late Mary Ritchie Key.

Even though the English words used for sense labels come with particular parts of speech (in English), the corresponding item in the described language may well have a different part of speech.

Naturally, not all concepts from groups (2) and (3) will be found in all the languages. In some cases, it may be important to add extra concepts to represent information relevant to a particular language group or a region.

At present (in February 2021) 333 IDS lists are available for online browsing and downloading at the main IDS website (see Figure 5).[17] The IDS is developed in cooperation and complementation with other research projects, and additional lists have been collected in such projects. The Loanword Typology project (LWT; Haspelmath & Tadmor 2009) has added 31 new languages out of a total of 41 languages investigated in the project on the basis of (somewhat extended) IDS lists,[18] and the Digital areal linguistics project described here together with a language documentation project also conducted in the same area have contributed an additional 16 languages, all from South Asia (generally also including the additional 150 LWT items),[19] plus an updated Swedish IDS list providing the linkage to SweFN++,[20] in the form of the standardized word sense identifiers used in SweFN++, and an English IDS list with Princeton WordNet synset identifiers (only open-class words).[21]

3.3.2 *The* IDS – *not just another Swadesh list*

The IDS was selected as the basis for our investigations partly because it provided a solidly established set of suitable concepts, but – more importantly – because we felt that the shorter Swadesh-style lists would be too small for the purpose of conducting a general comparative study of the language varieties spoken in the Himalayas.

We noted above that basic vocabulary lists tend to be short, especially those used in comparative studies, where an important aim is to cover a broad spectrum of languages. This has some practical and methodological consequences, to which we now turn.

---

17. https://ids.clld.org/

18. The database from the LWT project is available for browsing and downloading at https://wold.clld.org/. The LWT master list includes all IDS senses, but adds a further 150 senses – most of them under two new topic headings – taking the total up to 1,460 items subdivided into 24 categories.

19. https://spraakbanken.gu.se/en/projects/digital-areal-linguistics

20. https://spraakbanken.gu.se/en/resources/lwt

21. https://spraakbanken.gu.se/en/resources/lwt-pwn

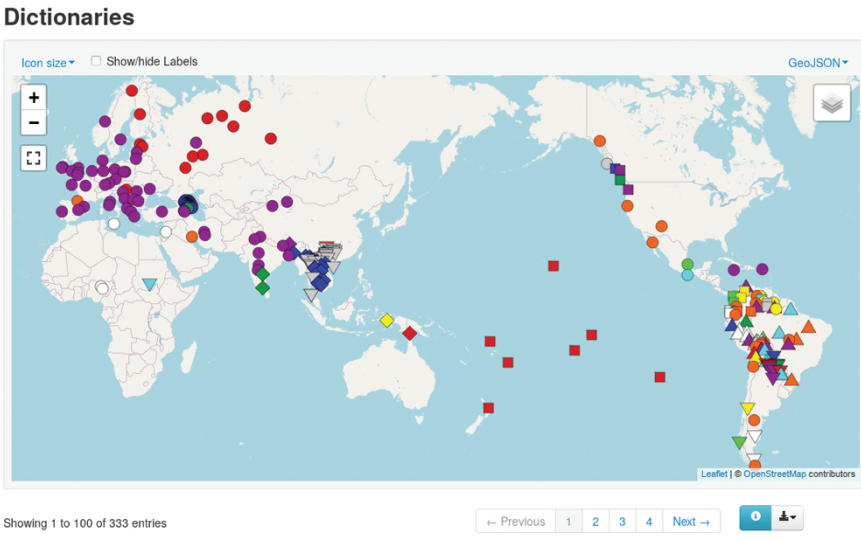


Figure 5. The geographical distribution of the (333) languages for which IDS lists are available

The incremental effort required to add a language to the database is small. Let us compare the IDS and ASJP in this regard. While the preparation of our IDS lists for South Asian languages has required on average one person-month of work per list, an ASJP list takes less than a day to prepare. The twenty-odd languages covered in Buck's original dictionary required about the same number of years from the start of the project (announced by Buck 1929) until the eventual publication of the dictionary (Buck 1949), but then Buck also included a wealth of detailed etymological material. From this follows that a project like ASJP can rely mainly on volunteer work, whereas this is much harder in the case of IDS. There is obviously a tension here between breadth and depth of coverage. The ASJP project, like much typological work (e.g., WALS: Dryer & Haspelmath 2013), aims for breadth at the expense of depth. It covers a significant fraction of the world's languages (9,788 lists covering 5,499 languages in version 19 of ASJP).

Size does matter, however. If the selection principles have been more or less similar, we can be fairly certain that the IDS list will be a much better basis for all kinds of linguistic investigations than the Swadesh lists, simply because it provides a broader empirical base. It is about an order of magnitude larger than the ASJP list, which is significant, since – as already noted (e.g. in Chapters 1 and 3) – many linguistic phenomena conform to a power-law (or long-tailed) distribution known as *Zipf's law* (Zipf 1949), one practical consequence of which is that data requirements increase exponentially as we wish to investigate increasingly rare linguistic

phenomena.[22] At the other end, is there some point below which language data just will not be useful for comparative purposes, when it will not be possible to say interesting (and true!) things about it? Apparently the ASJP list with its 28–40 items can be used to investigate many interesting aspects of genetic linguistics and language change, such as: the incidence of sound symbolism in certain basic vocabulary items (Wichmann, Holman & Brown 2010), the probable locations of language family homelands (Wichmann, Müller, et al. 2010), and the relationship between word length and rate of lexical change (Wichmann & Holman 2013). However, most details of the language systems remain hidden when using 28–40 central vocabulary items from each language, and nothing else. For instance, conducting systematic experiments with varying-size wordlists of Australian languages Dockum & Bowern (2018) find that at least 400 items are required for recovering the phoneme inventory of a language.

A relevant comparison in this connection could be the vocabulary needed for communicating in a foreign language. The ASJP vocabulary is on a par with that found in many tourist guide phrase lists, typically a column or two of words and short phrases, which everyone knows is sufficient only for displaying good intentions to native speakers, but hardly for any kind of real communication to be possible. The IDS in turn provides a vocabulary comparable in size to that defined for the CEFR A1 level, the lowest communicative proficiency level in a foreign language in the CEFR framework (around 1,500 items; Milton 2009: 186). At the A1 level, the learners should be capable of the following (COE 2012):

- Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type.
- Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has.
- Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

A basic premise of the IDS endeavor is that a vocabulary of a size which is capable of supporting linguistic activities at this quite impressive level of sophistication also will enable equally impressive broad comparative linguistic studies.

The first results of the study have been published, using reduced versions of the IDS lists (Saxena & Borin 2011, 2013) and a monograph on the linguistic situation of Kinnaur is under preparation by Anju Saxena, where full IDS lists are used in a number of comparative investigations of two Sino-Tibetan and one Indo-Aryan language local to the region (Saxena forthcoming).

---

22. At the next order of magnitude – about 64,000 items – we find full-size reference dictionaries which however are available only for a very small share of the world's languages.

4. Conclusion and future prospects

Areal and typological linguistics traditionally work with secondary language data, i.e. dictionaries and descriptive grammars. In the last few years, these disciplines have moved into the computer age, compiling large databases of selected linguistic features for many languages (see, e.g., Dryer & Haspelmath 2013; Everaert et al. 2009; Greenhill et al. 2008; Haspelmath & Tadmor 2009; Lewis & Xia 2010; Nerbonne 1998). The computer now gives us the potential for tying together these linguistic databases (see, e.g., Chiarcos et al. 2012), but not without a conscious effort. In this connection, SweFN++ makes a tangible contribution, ensuring that cross-linguistic lexical data are integrated into a well-defined formal structure both with respect to the adopted information model and the data formats used.

Funding

The research presented here has been supported by the European Commission (the *KELLY* project: no. 505630), by the Swedish Research Council (the projects *Swedish FrameNet++*: grant 2010–6013; *Digital areal linguistics*: grant 2009–01448; *Documentation of an endangered language: Kunashi*: grant 2014–00560; *South Asia as a linguistic area?*: grant 2014–00969), and by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken Text.

References

Ahlberg, Malin, Lars Borin, Markus Forsberg, Olof Olsson, Anne Schumacher & Jonatan Uppström. 2016. Karp: Språkbanken's open lexical infrastructure. In *Globalex 2016 book of abstracts*.

Baroni, Marco & Silvia Bernardini (eds.). 2006. *Wacky! Working papers on the Web as Corpus*. Online version: wackybook.sslmit.unibo.it. Bologna: GEDIT.

Berlin, Brent & Paul Kay. 1969. *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.

Borin, Lars. 2012. Core vocabulary: A useful but mystical concept in some kinds of linguistics. In Diana Santos, Krister Lindén & Wanjiku Ng'ang'a (eds.), *Shall we play the Festschrift game? Essays on the occasion of Lauri Carlson's 60th birthday*, 53–65. Berlin: Springer. https://doi.org/10.1007/978-3-642-30773-7_6

Borin, Lars, Bernard Comrie & Anju Saxena. 2013. The Intercontinental Dictionary Series – a rich and principled database for language comparison. In Lars Borin & Anju Saxena (eds.), *Approaches to measuring linguistic differences*, 285–302. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110305258.285

Borin, Lars, Markus Forsberg, Leif-Jöran Olsson & Jonatan Uppström. 2012. The open lexical infrastructure of Språkbanken. In *Proceedings of LREC 2012*, 3598–3602. Istanbul: ELRA.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, 474–478. Istanbul: ELRA.

Brysbaert, Marc & Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41(4): 977–990. https://doi.org/10.3758/BRM.41.4.977

Buck, Carl Darling. 1929. Words for world, earth and land, sun. *Language* 5(4): 215–227. https://doi.org/10.2307/409589

Buck, Carl Darling. 1949. *A dictionary of selected synonyms in the principal IndoEuropean languages*. Chicago: University of Chicago Press.

Campbell, Justice (Sir George). 1866a. Appendix A: List of words and phrases to be noted and used as test words for the discovery of the radical affinities of languages, and for easy comparison. In *Journal of the Asiatic Society. Special number: Ethnology*, vol. XXXV, p. II, 201–203. Calcutta: Asiatic Society of Bengal.

Campbell, Justice (Sir George). 1866b. The ethnology of India. In *Journal of the Asiatic Society. Special number: Ethnology*, vol. XXXV, p. II, 1–152. Calcutta: Asiatic Society of Bengal.

Chiarcos, Christian, Sebastian Nordhoff & Sebastian Hellman (eds.). 2012. *Linked data in linguistics: Representing and connecting language data and language metadata*. Berlin: Springer. https://doi.org/10.1007/978-3-642-28249-2

COE. 2012. *Common European Framework of Reference for languages: Learning, teaching, assessment (CEFR)*. Strasbourg: Council of Europe.

Dockum, Rikker & Claire Bowern. 2018. Swadesh lists are not long enough: Drawing phonological generalizations from limited data. In Peter K. Austin (ed.), *Language documentation and description*, vol 16, 35–54. London: EL Publishing.

Dolgopolsky, Aharon B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia. In Vitalij V. Shevoroshkin & Thomas L. Markey (eds.), *Typology, relationship and time: A collection of papers on language change and relationship by Soviet linguists*, 27–50. Ann Arbor: Karoma.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Jena: Max Planck Institute for the Science of Human History.

Eberhard, David M., Gary F. Simons & Charles D. Fennig (eds.). 2021. *Ethnologue: Languages of the world*. 24th edn. Dallas: SIL International.

Ebert, Karen. 2006. South Asia as a linguistic area. In Keith Brown (ed.), *Encyclopedia of languages and linguistics*, 2nd edn. Oxford: Elsevier. https://doi.org/10.1016/Bo-08-044854-2/00214-5

Emeneau, Murray. 1956. India as a linguistic area. *Language* 32: 3–16. https://doi.org/10.2307/410649

Evans, Nicholas. 2011. Semantic typology. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 504–533. Oxford: Oxford University Press.

Evans, Nicholas & Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32: 429–492. https://doi.org/10.1017/S0140525X0999094X

Everaert, Martin, Simon Musgrave & Alexis Dimitriadis (eds.). 2009. *The use of databases in cross-linguistic studies*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110198744

Fellbaum, Christiane. 1998a. Introduction. In Christiane Fellbaum (ed.), *WordNet: An electronic lexical database*, 1–19. Cambridge, Mass.: MIT Press. https://doi.org/10.7551/mitpress/7287.003.0004

Fellbaum, Christiane (ed.). 1998b. *WordNet: An electronic lexical database*. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/7287.001.0001

Forsbom, Eva. 2006. A Swedish base vocabulary pool. In *Proceedings of SLTC 2006*. https://cl.ling-fil.uu.se/~evafo/Papers/sltc06_forsbom_ea.pdf. Gothenburg: University of Gothenburg.

Georg, Stefan. 2017. Other isolated languages of Asia. In Lyle Campbell (ed.), *Language isolates*, 139–161. London: Routledge. https://doi.org/10.4324/9781315750026-6

Goddard, Cliff. 2001. Lexico-semantic universals: A critical overview. *Linguistic Typology* 5: 1–65. https://doi.org/10.1515/lity.5.1.1

Goddard, Cliff (ed.). 2008. *Cross-linguistic semantics*. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.102

Goddard, Cliff. 2012. Semantic primes, semantic molecules, semantic templates: Key concepts in the NSM approach to lexical typology. *Linguistics* 50(3): 711–743. https://doi.org/10.1515/ling-2012-0022

Greenhill, Simon J., Robert Blust & Russell D. Gray. 2008. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 2008(4): 271–283. https://doi.org/10.4137/EBO.S893

Grierson, George A. 1903–1927. *A linguistic survey of India*. Vol. I–XI. Calcutta: Government of India, Central Publication Branch.

Gustafson-Capková, Sofia & Britt Hartmann. 2006. *Manual of the Stockholm Umeå Corpus version 2.0*. Stockholm University, Dept. of Linguistics.

Hammarström, Harald, Robert Forkel & Martin Haspelmath (eds.). 2020. *Glottolog 4.3*. Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.4061162

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1): 31–80. https://doi.org/10.1515/flin.2011.002

Haspelmath, Martin & Uri Tadmor (eds.). 2009. *Loanwords in the world's languages: A comparative handbook*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110218442

Heine, Bernd & Tania Kuteva. 2005. *Language contact and grammatical change*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511614132

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller & Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica* 42(2): 331–354. https://doi.org/10.1515/FLIN.2008.331

Huang, Chu-ren, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari & Laurent Prevot (eds.). 2010. *Ontology and the lexicon: A natural language processing perspective*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511676536

Kay, Paul & Chad K. McDaniel. 1978. The linguistic significance of the meanings of basic color terms. *Language* 54(3): 610–646. https://doi.org/10.1353/lan.1978.0035

Kilgarriff, Adam, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi & Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation* 48: 121–163. https://doi.org/10.1007/s10579-013-9251-2

Koptjevskaja-Tamm, Maria. 2012. New directions in lexical typology. *Linguistics* 50(3): 373–394. https://doi.org/10.1515/ling-2012-0013

Kuhn, Tobias. 2014. A survey and classification of controlled natural languages. *Computational Linguistics* 40(1): 121–170. https://doi.org/10.1162/COLI_a_00168

Kuhn, Tobias. 2016. The controlled natural language of Randall Munroe's *Thing explainer*. In *Proceedings of CNL 2016*, 102–110. Cham: Springer. https://doi.org/10.1007/978-3-319-41498-0_10

Laming, Donald. 1957. *Human judgment: The eye of the beholder*. Andover: Cengage Learning.

Levinson, Stephen C. 2003. Language and mind: Let's get the issues straight! In Dedre Gentner & Susan Goldin-Meadow (eds.), *Language in mind: Advances in the study of language and thought*, 25–46. Cambridge, Mass.: MIT Press.

Lewis, William D. & Fei Xia. 2010. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing* 25(3): 303–319. https://doi.org/10.1093/llc/fqq006

Lyngfelt, Benjamin, Lars Borin, Kyoko Ohara & Tiago Timponi Torrent (eds.). 2018. *Constructicography: Constructicon development across languages*. Amsterdam: John Benjamins. https://doi.org/10.1075/cal.22

Malm, Per, Shafqat Virk, Lars Borin & Anju Saxena. 2018. LingFN: Towards a framenet for the linguistics domain. In *Proceedings of the International FrameNet workshop at LREC 2018: Multilingual framenets and constructicons*, 37–43. Miyazaki: ELRA.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard & David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL 2014*, 55–60. Baltimore: ACL. https://doi.org/10.3115/v1/P14-5010

Masica, Colin P. 1976. *Defining a linguistic area: South Asia*. Chicago: University of Chicago Press.

Matras, Yaron & Jeanette Sakel (eds.). 2007. *Grammatical borrowing in cross-linguistic perspective*. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110199192

Milton, James. 2009. *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters. https://doi.org/10.21832/9781847692092

Nerbonne, John (ed.). 1998. *Linguistic databases*. Palo Alto: CSLI, Stanford University.

Nichols, Johanna. 2003. Diversity and stability in language. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 283–310. Oxford: Blackwell. https://doi.org/10.1002/9780470756393.ch5

Ogden, Charles K. 1930. *Basic English: A general introduction with rules and grammar*. London: Paul Treber & Co., Ltd.

Oswalt, Robert L. 1971. Towards the construction of a standard lexicostatistic list. *Anthropological Linguistics* 13(9): 421–434.

Saxena, Anju. Forthcoming. *The linguistic landscape of the Indian Himalayas: Languages in Kinnaur*. Leiden: Brill.

Saxena, Anju & Lars Borin. 2011. Dialect classification in the Himalayas: A computational approach. In *Proceedings of Nodalida 2011*, 307–310. Riga: NEALT.

Saxena, Anju & Lars Borin. 2013. Carving Tibeto-Kanauri by its joints: Using basic vocabulary lists for genetic grouping of languages. In Lars Borin & Anju Saxena (eds.), *Approaches to measuring linguistic differences*, 175–198. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110305258.175

Slaska, Natalia. 2005. Lexicostatistics away from the armchair: Handling people, props and problems. *Transactions of the Philological Society* 103(2): 221–242. https://doi.org/10.1111/j.1467-968X.2005.00152.x

Swadesh, Morris. 1948. The time value of linguistic diversity. In *Paper presented at the Viking Fund supper conference for anthropologists*. Abstract in part: Swadesh 1952: 454.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16: 157–167. https://doi.org/10.1086/464084

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96(4): 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21: 121–137. https://doi.org/10.1086/464321

Tadmor, Uri. 2009. Loanwords in the world's languages: Findings and results. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 55–75. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110218442.55

Thomason, Sarah Grey. 2000. Linguistic areas and language history. In Dicky G. Gilbers, John Nerbonne & Jos Schaeken (eds.), *Languages in contact*, 311–327. Amsterdam: Rodopi.

Thomason, Sarah Grey & Terrence Kaufman. 1988. *Language contact, creolization and genetic linguistics*. Berkeley: University of California Press. https://doi.org/10.1525/9780520912793

Torrent, Tiago Timponi, Michael Ellsworth, Collin F. Baker & Ely Edison da Silva Matos. 2018. The Multilingual FrameNet shared annotation task: A preliminary report. In *Proceedings of the International FrameNet workshop 2018: Multilingual framenets and constructicons*, 62–68. Miyazaki: ELRA.

van der Auwera, Johan. 2012. From contrastive linguistics to linguistic typology. *Languages in Contrast* 12(1): 69–86. https://doi.org/10.1075/lic.12.1.05auw

Voice of America. 2009. *VOA Special English word book a–z*. Washington: Voice of America.

von Fintel, Kai & Lisa Matthewson. 2008. Universals in semantics. *The Linguistic Review* 25(1–2): 139–201. https://doi.org/10.1515/TLIR.2008.004

Wichmann, Søren & Eric W. Holman. 2013. Languages with longer words have more lexical change. In Lars Borin & Anju Saxena (eds.), *Approaches to measuring linguistic differences*, 249–281. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110305258.249

Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2010. Sound symbolism in basic vocabulary. *Entropy* 12(4): 844–858. https://doi.org/10.3390/e12040844

Wichmann, Søren, André Müller & Viveka Velupillai. 2010. Homelands of the world's language families: A quantitative approach. *Diachronica* 27(2): 247–276. https://doi.org/10.1075/dia.27.2.05wic

Wierzbicka, Anna. 1996. *Semantics: Primes and universals*. Oxford: Oxford University Press.

Wilks, Yorick. 2009. Ontotherapy, or how to stop worrying about what there is. In Nicolas Nicolov, Galia Angelova & Ruslan Mitkov (eds.), *Recent advances in natural language processing V*, 1–20. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.309.02wil

Xu, Hai. 2012. A critique of the controlled defining vocabulary in Longman Dictionary of Contemporary English. *Lexikos* 22: 367–381. https://doi.org/10.5788/22-1-1013

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.