LSK (Linguistic Society of Korea) 2022 winter conference, 2022 December 10

Comparing languages and comparing typological databases

Martin Haspelmath
Max Planck Institute for Evolutionary Anthropology (Leipzig)

1. Comparative databases of grammatical features

Since 2008, more and more online databases of grammatical parameters have become available.

The idea of studying grammatical variation systematically at a worldwide scale has existed for a century (Schmidt 1926; Greenberg 1963; Comrie 1989 etc.).

But for a long time, most grammarians have focused on the in-depth study of particular languages. Since the 1960s, many have pursued the idea of innate architectural universals, or of substantive universals – invariant elements of the innate universal grammar (UG). But this did not lead to large-scale comparative studies.

Since 1978, some authors of typological papers have given data tables with a few dozen languages, e.g. Ultan (1978); Stassen (1997) gives a table with 410 languages.

In the generative community, Cinque (1999) and Julien (2002) are two of the first works that listed dozens of languages; in general, works in the Chomskyan tradition have focused on "depth of analysis" rather than breadth of coverage.

In 2008, the online database of WALS came out (*World Atlas of Language Structures*, Dryer & Haspelmath 2008; 2011; 2013), based on the printed book (Haspelmath et al. 2005).

WALS was followed by a number of further databases that were published in the same framework ("CLLD", programmed by Robert Forkel):

APiCS (Atlas of Pidgin and Creole Language Structures, https://apics-online.info/)
SAILS (South American Indian Language Structures, https://sails.clld.org/)
eWAVE (Electronic World Atlas of Varieties of English)
PHOIBLE (segment inventory database, https://phoible.org/)
ValPaL (Valency Patterns Leipzig, https://valpal.info/)

There are now more and more other grammatical databases, created by research groups not associated with MPI-EVA, e.g.

SMG databases (Surrey Morphology Group, e.g. https://pips.surrey.ac.uk/)
DiaCL (Diachronic Atlas of Comparative Linguistics)
TALD (Typological Atlas of the Languages of Daghestan, http://lingconlab.ru/dagatlas/index.html)
SSWL (Syntactic Strucutures of the World's Languages, https://terraling.com/groups/7)

And very recently, a preliminary report on a forthcoming big grammatical database was published:

> Grambank (Skirgård et al. 2022): about 2400 languages, 195 features

2. Empirical challenges

Getting comparable information on the world's languages is difficult, because information on them is very unevenly described – the great majority of linguists work on the larger languages.

There is an increasing amount of information available on the world's languages – Hammarström et al. (2018) find that close to 25% of the world's languages have full grammars (more than 300 pages).

But extracting this information is very time-consuming, and for many questions, the grammars do not give us the answers that we want to ask (cf. Lesage et al. 2022).

Some other approaches:

> – **typological questionnaires**
> (e.g. https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaires.php, http://tulquest.huma-num.fr/taxonomy/term/44)
>
> – **parallel texts**
> (e.g. Cysouw & Wälchli 2007)

But these methods have downsides, too:

typological questionnaires:	who answers the questionnaire?
parallel texts:	the texts are unglossed...

3. Conceptual challenges (I): Finding the true natural categories for comparison

How do we find the right categories for comparison? Can we compare languages in terms of "subject" and "object" when we don't really know how to identify syntactic functions across languages?

> cf. Dryer (1997) on then non-universality of "subject" and "object"
> but Dryer (2005) looks at the order of "subject, object and verb"!!

Comparison of languages could (or should?) happen in terms of the **"natural parts"** that languages are made up of.

Linguists often presuppose that their categories and features are **natural kinds**, i.e. aspects of the innate language faculty (UG), e.g. distinctive features in phonology (Chomsky & Halle 1968), and:

> "We require that the grammar of a given language be constituted in accord with a specific theory of linguistic structure in which such terms as "phoneme" and "phrase" are defined independently of any particular language." (Chomsky 1957: 50)

A well-known example from Chomsky (1970):

[±N], [±V]:	noun:	[+N, –V]
	verb:	[–N, +V]
	adjective:	[+N, +V]
	adposition:	[–N, –V]

In other words, **universal grammar** provides a "toolbox" of categories that languages may use (Jackendoff 2002).

But what are the true **natural kinds** of language structure?

We do not really know, and linguistics has no clear criteria for assessing whether a feature or category should be assumed to be part of the innate language faculty (a natural kind) (cf. Haspelmath 2018, blogpost: https://dlc.hypotheses.org/1012).

The typical linguistics paper considers a narrow range of phenomena from a small number of languages and **provides an elegant account of the phenomena**, making use of some previously proposed general mechanisms and features.

It could be that this method will eventually lead to **convergent results**, and many linguists apparently have this hope, but I do not see much evidence for this over the last 50 years.

4. Conceptual challenges (II): Making sure that the comparisons are valid (comparative concepts)

If we want to make sure that the comparisons are valid, we must compare languages in terms of concepts which are **defined in the same way in all languages**.

Each language has its own unique structure (Haspelmath 2020), and its own unique categories – each language must be described in its own terms (Boas 1911).

So we cannot use language-particular descriptive categories to compare languages from around the world – we must use a special set of **comparative concepts** (Haspelmath 2010; 2018).

Consider the notion of "case":

in WALS, both **Baerman & Brown (2005)** and **Iggesen (2005)**
examine case marking in about two hundred languages,
with substantial overlap of languages

69 languages:	both B&B and I:	no case		
63 languages:	both B&B and I:	case		
7 languages:	B&B:	case	I:	no case
32 languages	B&B:	no case	I:	case

It appears that they use different ways of identifying "case" vs. "no case".

What about Japanese Korean "case particles" – are they postpositions or suffixes?

(1) Korean (Chae 2020: 133)
Wuli-nun siktang ***=eyse*** *achim pap* ***=ul*** *mek-ess-ta.*
we-CT restaurant =in morning meal =ACC eat-PST-DECL
'We ate breakfast in a restaurant.'

It depends on the definition of "case affix", and more specifically on the definition of "affix" vs. "clitic".

(2) **affix**
An affix is a bound morph that is not a root, that occurs on a root,
and that cannot occur on roots of different root classes. (Haspelmath 2021)

(3) **clitic**
A clitic is a bound morph that is neither an affix nor a root. (Haspelmath 2023a)

These definitions can be applied to all languages using the same criteria. The concepts "morph" (Haspelmath 2020b), "bound", "root", and "root class" (Haspelmath 2023b) apply to all languages in the same way.

By contrast, much of the earlier literature does not require using the same criterion in all languages – on the contrary, Zwicky (1985) explicitly says that different languages may show different "symptoms" of clitichood.

Linguistic categories are compared with **diseases**, where different patients may show different **symptoms** – but this makes sense only if linguistic categories are **natural kinds** (Haspelmath 2015; 2018).

So are Korean "case particles" affixes (in comparative terms)?

Nakamura (2018: 249): Japanese "case particles" may follow restrictive focus markers like 'only', which are clitics because they are not class-selective:

(4) *Hanako =dake =ga*
Hanako =only =NOM
'only Hanako (NOM)' (for Korean, see also Chae 2020: 39-40; 140)

Thus, Japanese "case particles" are not suffixes, which means they are postpositions.

(Baerman & Brown 2005: **no cases**; Iggesen 2005: **8-9 cases**)

Thus, valid comparison needs comparative concepts, defined uniformly for all languages.

5. Making comparative concepts commensurable across datasets

We want to make comparisons possble across databases, such as WALS, SAILS, SSWL, and so on.

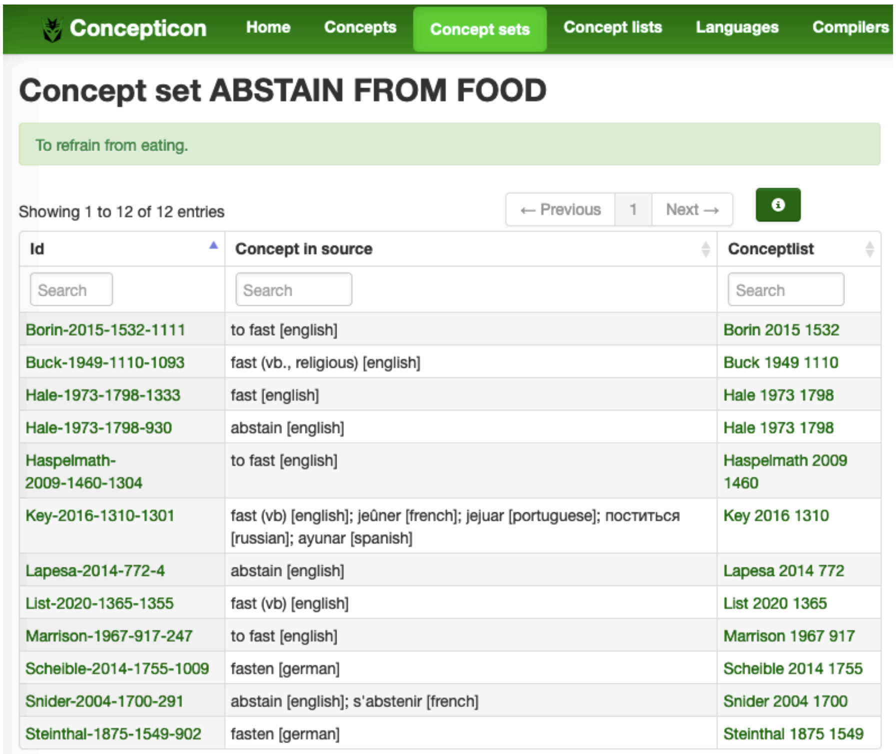
5.1. Comparisons of lexical databases: the Concepticon

The task of making grammatical databases comparable is similar to the task of **lexical comparison across languages** by means of a set of comparison meanings.

For lexical databases, a standard ontology now exists: The **Concepticon** (List et al. 2022, **concepticon.clld.org**), which has almost 4000 comparison meanings that bring together lexical concepts from diverse lexical data collections. This allows quick and automatic comparison of lexical forms from diverse databases.

The concepticon includes lexical concepts from 161 concept lists: Swadesh list, IDS list, SIL-Africa list, Sutton & Walsh Australian list, and so on.

Altogether, there are 116,000 lexical concepts, grouped together into about 4000 concept sets or metaconcepts.



Concept set ABSTAIN FROM FOOD

To refrain from eating.

Showing 1 to 12 of 12 entries

Id	Concept in source	Conceptlist
Borin-2015-1532-1111	to fast [english]	Borin 2015 1532
Buck-1949-1110-1093	fast (vb., religious) [english]	Buck 1949 1110
Hale-1973-1798-1333	fast [english]	Hale 1973 1798
Hale-1973-1798-930	abstain [english]	Hale 1973 1798
Haspelmath-2009-1460-1304	to fast [english]	Haspelmath 2009 1460
Key-2016-1310-1301	fast (vb) [english]; jeûner [french]; jejuar [portuguese]; поститься [russian]; ayunar [spanish]	Key 2016 1310
Lapesa-2014-772-4	abstain [english]	Lapesa 2014 772
List-2020-1365-1355	fast (vb) [english]	List 2020 1365
Marrison-1967-917-247	to fast [english]	Marrison 1967 917
Scheible-2014-1755-1009	fasten [german]	Scheible 2014 1755
Snider-2004-1700-291	abstain [english]; s'abstenir [french]	Snider 2004 1700
Steinthal-1875-1549-902	fasten [german]	Steinthal 1875 1549

5.2. A counterpart of of Concepticon: the Grammaticon

I am planning to set up a counterpart of the Concepticon for grammatical patterns, called **Grammaticon**, which will facilitate the comparison of different grammatical datasets.

Analogous to the lexical comparison meanings in the Concepticon (the **concept sets**), the Grammaticon contains **metafeatures** which capture what is common in highly similar features of different databases.

e.g.

WALS:	"Order of Subject, Object and Verb: SVO"
APiCS:	"Order of subject, object and verb: Subject-verb-object (SVO)"
SAILS:	"The dominant constituent order in a transitive clause is: AVP"
DiACL:	"What is the canonical (neutral) word order in a main clause? SVO"
SSWL:	"Property 05_SVO"

metafeature:

dominant order in transitive clauses is A-V-P

e.g.

WALS:	"Inclusive/exclusive distinction in independent pronouns: No inclusive/exclusive"
APiCS:	"Inclusive/exclusive distinction in independent personal pronouns: No inclusive/exclusive distinction"
SAILS:	"Is there an inclusive/exclusive distinction in personal pronouns? – no"
SAILS:	"Is there a distinction between inclusive and exclusive for personal pronouns? – no"

metafeature:

no clusivity distinction in independent personal pronouns

e.g.

Grambank:	"Can the recipient in a ditransitive construction be marked like the monotransitive patient? NO (0)"
WALS:	"Ditransitive Constructions: The Verb 'Give': Indirect-object construction"

metafeature:

the R-argument is not even partially aligned in coding with the P-argument

This requires setting up a set of grammatical terms with standard meanings (as comparative concepts) that can then be matched with the concepts used in the diverse databases.

Ultimately, one can perhaps hope that there will be a standard set of grammatical term that is widely known across the discipline – somewhat like the concepts used by projects in computational linguistics, e.g.

Universal Dependencies (de Marneffe et al. 2021)
UniMorph (Kirov et al. 2018)

References

Baerman, Matthew & Brown, Dunstan. 2005. Case syncretism. In Martin Haspelmath, David Gil & Bernard Comrie, Matthew S. Dryer (ed.), *The world atlas of language structures,* 118–121. Oxford: Oxford University Press.

Chae, Hee-Rahk. 2020. *Korean morphosyntax: Focusing on clitics and their roles in syntax*. Abingdon: Routledge.

Chomsky, Noam A. 1957. *Syntactic structures*. 's-Gravenhage: Mouton.

Chomsky, Noam A. 1970. Remarks on nominalization. In Jacobs, R.A. & Rosenbaum, Peter S. (eds.), *Readings in English transformational grammar*, 184–221. Waltham, MA: Ginn.

Comrie, Bernard. 1989. *Language universals and linguistic typology: Syntax and morphology*. Oxford: Blackwell.

Cysouw, Michael & Wälchli, Bernhard. 2007. Parallel texts: Using translational equivalents in linguistic typology. *STUF-Sprachtypologie und Universalienforschung* 60(2). 95–99.

de Marneffe, Marie-Catherine & Manning, Christopher D. & Nivre, Joakim & Zeman, Daniel. 2021. Universal Dependencies. Computational Linguistics 47(2). 255–308. (doi:10.1162/coli_a_00402)

Dryer, Matthew S. 1997. Are grammatical relations universal? In Bybee, Joan L. & Haiman, John & Thompson, Sandra A. (eds.), *Essays on language function and language type: Dedicated to T. Givón*, 115–143. Amsterdam: Benjamins.

Dryer, Matthew S. 2005. Order of subject, object and verb. In Haspelmath, Martin & Dryer, Matthew S. & Gil, David & Comrie, Bernard (eds.), *The World Atlas of Language Structures*, 330–333. Oxford: Oxford University Press. (https://wals.info/chapter/81)

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, Joseph H. (ed.), *Universals of language*, 73–113. Cambridge, MA: MIT Press.

Hammarström, Harald & Castermans, Thom & Forkel, Robert & Verbeek, Kevin & Westenberg, Michel A. & Speckmann, Bettina. 2018. Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation*. University of Hawaii Press 12. 359–392.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687. (doi:10.1353/lan.2010.0021)

Haspelmath, Martin. 2015. Defining vs. diagnosing linguistic categories: A case study of clitic phenomena. In Błaszczak, Joanna & Klimek-Jankowska, Dorota & Migdalski, Krzysztof (eds.), *How categorical are categories? New approaches to the old questions of noun, verb, and adjective*, 273–304. Berlin: De Gruyter Mouton.

Haspelmath, Martin. 2018. How comparative concepts and descriptive linguistic categories are different. In Van Olmen, Daniël & Mortelmans, Tanja & Brisard, Frank (eds.), *Aspects of linguistic variation: Studies in honor of Johan van der*

Auwera, 83–113. Berlin: De Gruyter Mouton. (https://zenodo.org/record/3519206)

Haspelmath, Martin. 2020a. The structural uniqueness of languages and the value of comparison for description. *Asian Languages and Linguistics* 1(2). 346–366. (doi:10.1075/alal.20032.has)

Haspelmath, Martin. 2020b. The morph as a minimal linguistic form. *Morphology* 30(2). 117–134. (doi:10.1007/s11525-020-09355-5)

Haspelmath, Martin. 2021. Bound forms, welded forms, and affixes: Basic concepts for morphological comparison. *Voprosy Jazykoznanija* 2021(1). 7–28. (doi:10.31857/0373-658X.2021.1.7-28)

Haspelmath, Martin. 2023a. Types of clitics in the world's languages. *(in preparation)*.

Haspelmath, Martin. 2023b. Word class universals and language-particular analysis. In van Lier, Eva (ed.), *Oxford handbook of word classes*. Oxford: Oxford University Press (to appear).

Himmelmann, Nikolaus P. 2022. Against trivializing language description (and comparison). *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 46(1). 133–160. (doi:10.1075/sl.19090.him)

Iggesen, Oliver A. 2005. Number of cases. In Haspelmath, Martin & Dryer, Matthew S. & Gil, David & Comrie, Bernard (eds.), *The world atlas of language structures*, 202–205. Oxford: Oxford University Press. (http://wals.info/chapter/49)

Kirov, Christo & Cotterell, Ryan & Sylak-Glassman, John & Walther, Géraldine & Vylomova, Ekaterina & Xia, Patrick & Faruqui, Manaal et al. 2020. UniMorph 2.0: Universal Morphology. arXiv. (doi:10.48550/arXiv.1810.11101) (http://arxiv.org/abs/1810.11101)

Lesage, Jakob & Haynie, Hannah J. & Skirgård, Hedvig & Weber, Tobias & Witzlack-Makarevich, Alena. 2022. Overlooked data in typological databases: What Grambank teaches us about gaps in grammars. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2884–2890. Marseille, France: European Language Resources Association. (https://aclanthology.org/2022.lrec-1.309) (Accessed December 9, 2022.)

List, Johann Mattis & Tjuka, Annika & Rzymski, Christoph & Greenhill, Simon & Forkel, Robert (eds.). 2022. *Concepticon 3.0.0.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://concepticon.clld.org/) (Accessed December 9, 2022.)

Schmidt, Wilhelm. 1926. *Die Sprachfamilien und Sprachenkreise der Erde*. Heidelberg: Winter.

Skirgård, Hedvig et al. 2022. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. SocArXiv. (https://osf.io/preprints/socarxiv/grh45/) (Accessed December 9, 2022.)

Zwicky, Arnold M. 1985. Clitics and particles. *Language* 61(2). 283–305.