ERC Advanced Grant 2014/15

Grammatical Universals:

"Form-frequency correspondences in grammar"

Proposal acronym:

FormGram

Principal Investigator:

MARTIN HASPELMATH

Host institution:

LEIPZIG UNIVERSITY (Universität Leipzig)

Duration: 60 months

Proposal summary

This project will document and explain a substantial number of grammatical universals by demonstrating a link between cross-linguistic patterns of language form and general trends of language use. The claim is that frequently expressed meanings tend to be expressed by short forms, not only at the level of words, but also throughout the grammars of languages around the world (**form-frequency correspondences**). A simple example is the asymmetry in the coding of present-tense forms and future-tense forms in the world's languages, as one out of a multitude of analogous cases: Present-tense forms tend to be short or zero-coded, while future-tense forms tend to be longer or to have an overt marker. This corresponds to a usage asymmetry: Present-tense forms are generally more frequent than future-tense forms, in all languages. The proposed explanation is that higher-frequency items are more predictable than lower-frequency items, and predictable content need not be expressed overtly or can be expressed by shorter forms. Form-frequency correspondences thus make language structure more efficient, but it still needs to be shown that there exists a mechanism that creates and maintains these efficient structures: recurrent instances of language change driven by the speakers' preference for user-friendly utterances. The project thus combines cross-linguistic research on grammar, cross-linguistic corpus research and historical linguistics in a ground-breaking way. For reasons that have to do with the history of the discipline, form-frequency correspondences are still largely overlooked and ignored by linguists, so the current project will have a significant impact on our general understanding of human language.

Extended Synopsis

Objectives

Universals of language structure are a key component in our understanding of the nature of human language. This project makes an important contribution to the documentation and explanation of grammatical universals by an extensive demonstration of links between cross-linguistic patterns of language form and general tendencies of language use. The prevalent view among theoretically oriented grammar researchers over the last few decades has been that key aspects of grammatical knowledge must be innate ("universal grammar"), and the primary question asked in this approach has been how the diverse grammatical patterns can be acquired by learners (e.g. Chomsky 2000). Language use, in particular frequency of use, plays no role in this approach, and the focus is not on explaining language form, but on explaining language acquisition.

In this project, a different tack is taken that has a potentially enormous impact on the way linguists understand grammar: I first ask what we can learn about the limits of variation in language form from existing grammatical descriptions, especially from the multitude of recently published excellent grammars of small, often endangered languages. I focus on a set of seven core areas of grammatical structure (listed below) and ask for each of them which cross-linguistic generalizations can be found.

(1)
1. Number and mass nouns
2. Possessive constructions
3. Locational marking
4. Adjectives vs. nouns and verbs
5. Tense and aspect
6. Valency-changing alternations
7. Cross-clausal coreference and disreference

Languages vary enormously at all levels of structure, and it has even been argued recently that language universals are a "myth" (Evans & Levinson 2009). However, if one's goal is not to discover a common blueprint for all language-specific grammars, but merely to identify pervasive general tendencies, the search is very often successful. In particular, we often find form asymmetries such that one kind of meaning is not overtly expressed ("expressed by zero", e.g. singular number, nominative case, present tense, same-subject complement), whereas a contrasting kind of meaning is expressed by an overt marker (e.g. plural number, locative case, future tense, different-subject complement). Some examples from different languages are given in (2)-(5) (the symbol "Ø" is used for zero, i.e. absence of a marker).

(2)		singular	plural
	Lezgian	*tar-Ø*	*tar-**ar***
		'tree'	'trees'
	Hup	*hup-Ø*	*hup-**d'əh***
		'person'	'persons'

(3)		nominative	locative
	Basque	*etxe-Ø*	*etxe-**an***
		'house'	'in the house'
	Hebrew	*Ø-beitenu*	***bə**-veitenu*
		'our house'	'in our house'

(4)		present tense	future tense
	Latin	*vide-Ø-t*	*vide-**bi**-t*
		'sees'	'will see'
	Yoruba	*nwọn Ø wá*	*nwọn **ó** wá*
		'they come'	'they will come'

(5)		same-subject	different-subject
	German	*will Ø sterben*	*will,* ***dass*** *er stirbt*
		'wants to die'	'wants him to die'
	Maltese	*jrid Ø jiġi*	*jrid* ***li*** *jiġi*
		'wants to come'	'wants him to come'

I propose that all these zero-overt contrasts can be explained by a simple efficiency consideration: Frequently expressed meanings tend to be expressed by short forms, not only at the level of words, but also throughout the grammars of languages around the world. This is what I call **form-frequency correspondences**: It can be shown that in general, singulars are more frequent than plurals, nominatives are more frequent than accusatives, and so on. The ultimate explanation is the **predictability** of frequent forms: Speakers can afford to expend less energy on predictable meanings. That language form can be related to language use has long been known (e.g. Zipf 1935, who mostly limited his attention to the length of words), and usage-based approaches to language have become more prominent recently, but the pervasiveness of form-frequency correspondences has barely been recognized by the field of linguistics.

This project is **innovative** in that it considers both cross-linguistic formal tendencies and cross-linguistic usage tendencies. While cross-linguistic structural patterns have been studied systematically for quite a while (cf. Haspelmath et al.'s (2005) *World Atlas of Language Structures*), cross-linguistic usage patterns have hardly been studied at all. In the past, this was difficult, but with the availability of more and better corpora from a wide range of languages, this is becoming feasible.

However, just noting that short forms for frequently expressed meanings yield economical, efficient systems is not sufficient as an explanation. In addition, we need a general mechanism that creates and maintains efficient language systems and serves as a link between the usage trends and the formal properties of languages. The proposed mechanism is **usage-based language change**, i.e. language change that is driven by the speakers' preference for user-friendly utterances (as simple as possible, but sufficiently clear for the hearer), and thus ultimately user-friendly structures. Like the form asymmetries and the frequency asymmetries, the diachronic trends that I would make responsible for the changes have to be highly general, and thus in principle they should be found throughout the world. Demonstrating this will be more difficult, but some important steps toward this goal are within the scope of the present project.

The **risk** of this project is that in a number of grammatical domains, it will be difficult to conclusively prove the effect of asymmetries of usage frequency, for a variety of reasons: (i) Frequency effects on form may be masked by interfering factors, especially if the frequency differences are small and hence the effect is not very strong; (ii) Some form-frequency correspondences may be visible only in a relatively small number of languages, so that it may not be easy to demonstrate a universal tendency for form asymmetry; (iii) Usage frequency asymmetries may be difficult to ascertain because of bias in our corpora (for example, this affects the frequency of different person forms, as face-to-face conversation is not widely represented in available corpora, so that first and second person forms are usually underrepresented). Moreover, the planned collaboration between historical linguistics, corpus linguistics and world-wide comparative linguistics presents a challenge for the project. However, the **gains** of the project, if successful, will be enormous: Linguists will have a much deeper understanding of how usage tendencies give rise to many of the most salient grammatical patterns in a wide variety of grammatical domains.

Project team

In addition to myself, the project team will include **four postdocs**, covering the three areas: world-wide structural comparison of languages, comparative corpus linguistics, and typologically oriented diachronic linguistics. They will work for five years on form-frequency correspondences in the seven subprojects identified above. I also request funding for three **student assistants**, in order to involve students in the research and to help us with easier tasks.

I also request funding for **short-term visiting scientists** (for 2-6-week stays, about six visitors per year, only travel and additional expenses covered), because serious interaction and collaboration with scientists at other institutions requires more time than short meetings at conferences.

State of the art

Grammatical universals have been studied systematically by some scholars for the last four decades or so, and we have quite a few proposed universals and universal trends (cf. Plank (ed.) 2002ff.), as well as more comprehensive documentation for over a hundred grammatical features (Haspelmath et al. 2005), but this has really scratched only the surface of the grammatical regularities. Many plausible hypotheses have not been verified by systematic investigations. The empirical basis for cross-linguistic research of the required type ("grammar mining") is becoming better and better due to the recent progress in grammar writing, and database technology is making the organization and dissemination of research results easier. The universal patterns that are of primary interest here have sometimes been studied under the rubric of "(typological) markedness" (Greenberg 1966, Croft 1990/2003), but this has not been a major focus of research more recently, and many form-frequency correspondences have been overlooked. Form asymmetries of the type studied here have often been explained as being due to "iconicity" or "cognitive simplicity" (Clark & Clark 1978, Haiman 1983, Bybee 2011), but frequency and predictability are superior as explanations (Haspelmath 2008a) because they also account for "markedness reversals", as when some nouns are exceptionally longer in the singular, or some nouns exceptionally have a zero locative form.

Usage-based research that reckons with effects of language use on language structure has become more and more widespread recently (e.g. Bod 2010, Bybee 2010, Jaeger & Tily 2011, Taylor 2012), but most of this research deals with the mental representation of grammatical knowledge and with language acquisition, rather than with the question why language structures are the way they are. Moreover, most current usage-based research is concerned primarily with the major languages. Cross-linguistic corpus-based studies are still in their infancy (works such as Wälchli 2009 being pioneering exceptions). But this is what theoretical linguists must aim for if they want to avoid errors deriving from the concentration on a few well-known languages whose properties do not generalize to the full diversity of human languages (cf. Evans & Levinson 2009).

To answer the question why languages are the way they are, we must be able to say how usage preferences become grammatical **conventions**, and we are only beginning to understand this. The most ambitious work in this area, Hawkins (2004) (where many form-usage correspondences are identified), has little to say on this question. But it is clear that the way in which usage preferences turn into conventions of grammar is by **diachronic change** (Bybee 1988, Haspelmath 2008b), so we need to study these pathways of change in order to provide a complete explanation of the observed trends.

Methodology

The general methodology is easy to describe: we need to establish general patterns of language structure, language use, and diachronic change across a significant number of diverse grammatical domains.

(i) We study the **structures** of many different languages from around the world on the basis of descriptive grammars and more specialized literature ("grammar mining") to establish and document universal trends for the domains specified (1). The methodology here is not new (this was Greenberg's method when he established his universals in the early 1960s), but due to the enormous progress of grammar writing over the last 15 years (largely as a result of the language documentation movement), a lot more and better information is available for mining than was available even quite recently.

(ii) We study the **usage frequency** of the relevant grammatical patterns in a sample of corpora from diverse languages, in order to determine universal frequency asymmetries of grammatical meanings. This is a very new type of research (but see Greenberg 1966, Croft 1991 for precursors), and since it is not yet easy to access larger corpora from many different languages simultaneously, we will have to work with compromises of various sorts. One aspect that needs to be kept in mind is that the corpus evidence is particularly convincing if it comes from languages that do not have overt coding asymmetries for the relevant semantic contrast (e.g. which have both an overt singular and an overt plural marker), because a skeptic could object that the frequency asymmetry is actually caused by the coding asymmetry (longer forms might be used more rarely simply because they are longer, cf. Haspelmath 2008b: 212).

(iii) We study **diachronic paths of change** that lead to the patterns under investigation. While corpus research presents many opportunities (apart from methodological challenges), comparative diachronic research presents far fewer opportunities. For the vast majority of the world's languages, we have no attested data going back more than a few decades, and the diachronic time scale that interests us is much deeper than that. Thus, we will have to rely on (i) those few languages (mostly European, Western Asian, East Asian, plus Egyptian) that do have a long attestation and whose history has been studied thoroughly, and (ii) reconstructed language histories. The latter exist for quite a few additional language families and are often quite reliable. Thus, our goals for the diachronic part of the project will have to be somewhat more modest than for the structure and usage parts.

Subprojects

Form-frequency correspondences can be found in many different areas of grammar, and it is the task of this project to study a representative set of them in greater depth. Some of these are well-known cases of "markedness", others have hardly been studied under this perspective:

– **Number and mass nouns:** Singular-plural pairs are a well-known example of a zero-overt contrast, but there are still many open issues, and mass nouns are rarely considered with them.

– **Possessive constructions:** Alienable/inalienable contrasts have usually been considered as explained by iconicity (e.g. Haiman 1983), but there is a clear frequency dimension here, and preliminary work indicates that the frequency account explains more patterns than iconicity (cf. Haspelmath 2008a: 19-22).

– **Locational marking:** Locative marking is generally longer than nominative and accusative marking because locatives are rarer. But sometimes locational NPs are zero-coded, apparently precisely when the locative use is more frequent than nominative use.

– **Adjectives vs. nouns and verbs:** Some semantic types of adjectives tend to be used cross-linguistically in a noun-like way, with copula in predicative function, while others are used in a verb-like way (Dixon 1977). This may be explained by the different frequencies in attributive and predicative position.

– **Tense and aspect:** For example, stative verbs tend to have imperfective present-tense meanings when unmarked, while dynamic verbs tend to have perfective past meanings when unmarked. It remains to be shown that this is truly a cross-linguistic trend, and it is very likely that a frequency correlate will be found.

– **Valency-changing alternations:** Passives, causatives and applicatives are less frequent than the active forms, and as a result they generally tend to be coded overtly. However, sometimes the causative is an unmarked form, and the anticausative is marked, providing crucial evidence for the frequency account. Similarly, languages with reciprocal markers often have some verbs that lack reciprocal marking.

– **Cross-clausal coreference and disreference:** Coreference is unusual within a clause, making reflexives rare and hence more heavily coded (Haspelmath 2008c), but cross-clausally, coreference is more usual, leading to shorter marking of coreferential pronouns (long-distance reflexives) and same-subject verb forms.

Theory

Although the fundamental link between frequency, predictability and shortness of grammatical marking is clear and can hardly be doubted, there are a number of general issues that need to be discussed in this project and by the discipline of linguistics as a whole.

– Ultimate cause: Could the correlation be explained by some other factor that causes both the frequency results and the formal patterns? This seems unlikely, because the frequency asymmetries may have quite diverse causes (cognitive preference, frequency in the world, occasions for use of a form, etc.), while the outcomes are quite uniform, but we should keep looking.

– Frequency vs. predictability: Is there a way to measure predictability of grammatical meanings that is independent of frequency of use? If so, is predictability a better predictor of shortness of coding?

– Instead of looking at grammatical forms and determining their frequencies, can we start out from corpora, identify frequency asymmetries, and then go on to ask whether these are reflected in form asymmetries?

– Sometimes the length contrast is between short and long markers, and sometimes between zero and overt markers. Is there anything that makes these two types different?

– To what extent is it helpful to model diachronic changes and synchronic languages states by means of formalisms or technical frameworks such as stochastic optimality theory (e.g. Bresnan et al. 2001) or evolutionary game theory (e.g. Jäger 2007)?

– A question that we will have in mind all the time is what the competing models are that could provide alternative accounts of some of our phenomena. In particular, we will constantly compare our predictions and the explanatory force of our theory with approaches based on Universal Grammar.

Deliverables

The **deliverables** are (i) research papers, (ii) a summarizing monograph, and (iii) published typological databases that are the outcome of the cross-linguistic research on language structures.

I envisage at least 1.5 published **papers** per year per scientist, so overall at least 30 papers, of which at least half (preferably more) should appear in top journals. The **summarizing monograph** will not repeat every finding from every paper of ours, but will summarize the main results of the project in a succinct way.

The **cross-linguistic databases** will be put together from the start with a view to publication, probably in the by now well-established CLLD framework (clld.org). High-quality cross-linguistic data is difficult to obtain, so we want to give full access to the results of our grammar mining to subsequent researchers.

References

Bod, Rens. 2010. Probabilistic linguistics. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press.

Bresnan, Joan, Shipra Dingare & Christopher D. Manning. 2001. Soft constraints mirror hard constraints: voice and person in English and Lummi. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG 01 Conference*, The University of Hong Kong, 13–32. CSLI Online.

Bybee, Joan L. 1988. The diachronic dimension in explanation. In J. A. Hawkins (ed.), *Explaining language universals*, 350–379. Oxford: Blackwell.

Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, Joan. 2011. Markedness: iconicity, economy, and frequency. In Jae Jung Song (ed.), *The Oxford Handbook of Linguistic Typology*, 131–147. Oxford: Oxford University Press.

Chomsky, Noam. 2000. *New horizons in the study of language and mind*. Cambridge: Cambridge University Press.

Clark, Eve V. & Herbert H. Clark. 1978. Universals, relativity, and language processing. In Joseph H Greenberg (ed.), *Universals of human language*, vol. 1: Method and theory, 225–277. Stanford: Stanford University Press.

Croft, William. 1990/2003. *Typology and universals.* (1st/2nd ed.) Cambridge: Cambridge University Press.

Croft, William. 1991. *Syntactic categories and grammatical relations: the cognitive organization of information.* Chicago: U. of Chicago Press.

Dixon, R.M.W. 1977. Where have all the adjectives gone? *Studies in Language* 1.1: 1-80.

Evans, Nicholas & Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32(05). 429–448.

Greenberg, Joseph H. 1966. *Language universals, with special reference to feature hierarchies.* The Hague: Mouton.

Haiman, John. 1983. Iconic and economic motivation. *Language* 59(4). 781–819.

Haspelmath, Martin. 2008a. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19.1:1-33.

Haspelmath, Martin. 2008b. Creating economical patterns in language change. In: Good, Jeff (ed.) 2008. *Linguistic universals and language change.* Oxford: Oxford University Press. 185-214.

Haspelmath, Martin. 2008c. A frequentist explanation of some universals of reflexive marking. *Linguistic Discovery* 6.1:40-63.

Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie (eds.) 2005. *The World Atlas of Language Structures.* Oxford: Oxford University Press.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Jaeger, T. Florian & Harry Tily. 2011. On language “utility”: Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(3). 323–335.

Jäger, Gerhard. 2007. Evolutionary Game Theory and typology: a case study. *Language* 83.1: 74-109.

Plank, Frans (ed.) 2002ff. *The universals archive.* http://typo.uni-konstanz.de/archive/

Taylor, John R. 2012. *The mental corpus : how language is represented in the mind*. Oxford: Oxford University Press.

Wälchli, Bernhard. 2009. *Motion events in parallel texts: a study in primary-data typology.* Bern: Universität Bern habilitationsschrift.

Zipf, George K. 1935. *The psycho-biology of language: an introduction to dynamic philology.* Boston: Houghton Mifflin.