J. Linguistics **57** (2021), 605–633.
doi:10.1017/S0022226720000535

Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability[1]

MARTIN HASPELMATH

Max Planck Institute for Evolutionary Anthropology
&
Leipzig University

(Received 21 December 2019; revised 17 July 2020)

This paper claims that a wide variety of grammatical coding asymmetries can be explained as adaptations to the language users' needs, in terms of frequency of use, predictability and coding efficiency. I claim that all grammatical oppositions involving a minimal meaning difference and a significant frequency difference are reflected in a universal coding asymmetry, i.e. a cross-linguistic pattern in which the less frequent member of the opposition gets special coding, unless the coding is uniformly explicit or uniformly zero. I give 25 examples of pairs of construction types, from a substantial range of grammatical domains. For some of them, the existing evidence from the world's languages and from corpus counts is already strong, while for others, I know of no counterevidence and I make readily testable claims. I also discuss how the functional-adaptive forces operate in language change, and I discuss a number of possible alternative explanations.

KEYWORDS: asymmetric coding, markedness, morphological universals, predictability, token frequency

1. OVERVIEW

In this paper, I formulate an explanation of universal grammatical coding asymmetries in terms of predictability and efficiency of coding. I claim that many of the widespread and systematic coding patterns that we find in the world's languages are due to well-understood functional-adaptive forces, and that there is no need to resort to 'markedness' or other poorly understood or idiosyncratic mechanisms.

By grammatical coding asymmetries, I mean minimal grammatical oppositions such as those in Table 1, where one member is typically zero-coded (or shorter),

[1] The support of the European Research Council (ERC Advanced Grant 670985, Grammatical Universals) is gratefully acknowledged. I also thank the editors and several reviewers, both for *Glossa* and the *Journal of Linguistics*, as well as my colleagues in the ERC project group (especially Katarzyna Janic, Natalia Levshina, Susanne Maria Michaelis, Karsten Schmidtke-Bode, Ilja Seržant), for help of various kinds.

singular	plural	(*book* – *book*-***s***)	§3.1
nominative (A/S)	accusative (P)	(*he* – *hi*-***m***)	§3.2
allative	ablative	(*to* – ***from***)	§3.3
positive	comparative	(*small* – *small*-***er***)	§3.6
present	future	(*go* – ***will*** *go*)	§4.1
affirmative	negative	(*go* – ***don't*** *go*)	§4.5
inanimate patient	animate patient	(Spanish Ø *la casa* – ***a*** *la mujer*)	§6.1
3rd person	2nd person	(Spanish *canta*$_{\text{3SG}}$ /*canta*-***s***$_{\text{2SG}}$ ‘sing(s)’)	§7.2
2nd person imp.	3rd person imp.	(*praise!* – ***let*** *her praise!*)	§7.2
attributive adjective	attributive verb	(*small* – *play*-***ing***)	§7.3

A = transitive subject, S = intransitive subject, P = direct object, Ø = no marker

Table 1

Examples of universal grammatical coding asymmetries.

while the other member has an overt coding element or contains more segments. The table shows only a single example for each opposition type, but the claim is that all these asymmetries are universal tendencies in the world's languages. The table's last column indicates which section of the present article deals with a given opposition.

The basic observation is that the zero-coded (or shorter) member of such an opposition is generally more frequent in language use. For example, singular nouns are more frequent than plural nouns, and present-tense forms are more frequent than future-tense forms. This can be related to the general observation in (1).

(1) *The form–frequency correspondence universal*
Languages tend to have shorter forms for more frequent meanings.

This is an old insight which is uncontroversial for word length (e.g. Zipf 1935: 23), but my claim here is that it is also generally valid for grammatical patterns, as formulated in (2) (which is just a special instance of (1)). The idea of form–frequency correspondences in grammar can be traced back to Greenberg (1966) (see also Croft 2003: Chapter 4; Hawkins 2004: Section 3.2.2; Haspelmath 2008a, b; Diessel 2019: Chapter 11).

(2) *The grammatical form–frequency correspondence hypothesis*
When two grammatical construction types that differ minimally (i.e. that form a semantic opposition) occur with significantly different frequencies, the less frequent construction tends to be overtly coded (or coded with more segments), while the more frequent construction tends to be zero-coded (or coded with fewer segments), if the coding is asymmetric.

Since the coding asymmetries illustrated in Table 1 are strong universal tendencies (sometimes even exceptionless), there must be a highly general explanatory factor. I claim (in line with the earlier work by Greenberg and others) that the explanation lies in the predictability of frequently used grammatical construction types. The causal chain is thus as in (3).[2]

(3) frequency of use → predictability → shortness of coding

This is thus a kind of 'economy' explanation (Haiman 1983): Speakers, and hence language systems, favour economical patterns, which require a greater amount of coding energy only for less predictable parts of linguistic messages. Hearers are more surprised by less frequent aspects of messages and thus need more robust coding for them. Another way of putting it is that language systems favour efficient coding (Hawkins 2004, 2014), or that they support efficient communication (Gibson et al. 2019).

This explanation implies that language systems have adapted to the users' needs. If language systems are viewed strictly from a synchronic and static perspective, as is still often the case in theoretical linguistics, this may appear puzzling. But language systems are flexible or malleable, in the sense that system-external functional-adaptive forces can have minute effects on the behaviour of individual speakers in individual utterances, which eventually have the cumulative result of creating efficient systems (Keller 1994). We can get a better sense of these processes when we consider the diachronic origins of grammatical coding asymmetries (see Section 8 below).

I thus defend a functional-adaptive explanation of universal coding asymmetries, but it should be noted that this is not a traditional functional explanation of the type that establishes a direct link between meaning and form (note Bybee's 1985 subtitle: 'A study of the relation between meaning and form'). On the contrary, I reject explanations of coding asymmetries in terms of iconicity (meaning–form matching) (see Haspelmath 2008a), and I will summarize the arguments briefly below (Section 9.1). Likewise, markedness cannot be an explanatory factor (as argued by Haspelmath 2006). Instead of linking meanings and shapes directly, the explanation that I defend here makes crucial reference to predictability.

I will say more about the causal chain in Section 9 below, but first I will explain the nature of the universal claims (Section 2), and then I will give 25 examples of universally asymmetric pairs of grammatical construction types. First, there are 13 examples of simple meaning pairs (such as singular/plural and present/future), both from the nominal domain (Section 3) and from the predicational domain (Section 4). These are followed by 12 differential-coding pairs in Sections 5–7. DIFFERENTIAL CODING is a term that I use for a situation where a grammatical meaning is coded in two distinct ways depending on the grammatical or lexical context, as in differential object marking in Spanish (*Veo la casa* 'I see the house' vs. *Veo **a** la mujer* 'I see the woman', see Section 5). Such situations are quite parallel to simple pairs of grammatical meanings, and I subsume both under the general notion of

[2] Predictability derives not only from frequency of use but also from presence in the context (see Section 9.2). The claim here is merely that frequency of use contributes to predictability in such a way that it influences grammatical coding.

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

'(pairs of) construction types'. I give six examples of differential-coding pairs from the nominal domain in Section 6, and six examples from the predicational domain in Section 7. Then in Section 8, I consider a number of paths of change that result in asymmetric coding patterns, and in Section 9, I discuss the nature of the predictability- and efficiency-based explanation in some more detail.

2. Universal grammatical coding asymmetries

In this paper, I do not have space to justify the universality of all the grammatical coding asymmetries that I discuss and that I claim fall under the scope of the explanatory theory. So I refer to the existing literature, which begins with Greenberg (1963), who formulated the universals in (4).

(4) (a) *Greenberg's Universal 35*
There is no language in which the plural does not have some nonzero allomorphs, whereas there are languages in which the singular is expressed only by zero. The dual and the trial are almost never expressed only by zero.

(b) *Greenberg's Universal 38*
Where there is a case system, the only case which ever has only zero allomorphs is the one which includes among its meanings that of the subject of the intransitive verb.

Many more universals were stated or implicit in Greenberg's later book (1966), where he noted that grammatical number, gender, person, case, tense, aspect, mood and other domains often show asymmetries which he described with the terms 'marked/unmarked'. He did not provide much documentation for the claims (any more than he did in his 1963 article), but they were based on his very wide-ranging knowledge of the world's languages, and it appears that none of his claims have been shown wrong.

In general, a universal coding asymmetry involving construction type 1 (e.g. singular) and construction type 2 (e.g. plural) implies the universal claim in (5).

(5) If a language has an asymmetric coding contrast between construction type 1 (universally more frequent) and construction type 2 (universally less frequent), then construction type 1 shows a strong tendency to be coded with a shorter shape than construction type 2, and often by zero.

Thus, the eleven pairs in Table 1 above stand for eleven implicational universals of the form in (5), and each of the 25 asymmetric pairs in Sections 3–7 corresponds to a universal.

Of course, when I claim that a grammatical coding asymmetry is universal, I do not mean to exclude the possibility of individual exceptions, and I merely claim that the asymmetry is found with much greater than chance frequency. This is conveyed by 'strong tendency' in (5).

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

It should be noted that the grammatical construction types that figure in the universals are comparative concepts, in the sense of Haspelmath (2010). Thus, there is no claim that the construction pairs are completely identical across languages, or that they must instantiate the same innate grammatical features or categories. If one wanted to explain these universals by biocognitive constraints imposed by an innate grammar blueprint, then one would expect languages to show the same categories and features, but for the functional-adaptive explanation that I propose here, all we need is comparability through clearly defined comparative concepts. For example, languages may have fairly different kinds of passives, and fairly different kinds of ablatives. But if a construction turns the P-argument of the active into the new S-argument, and the A-argument is oblique-marked and optional or absent, then it counts as a passive for the purposes of the comparison. And if a case-marker or adposition expresses a spatial source, then it counts as an ablative marker, whatever its other properties.

Now of course, not all languages and all grammatical contrasts show asymmetric coding. For most of the universal coding asymmetries, there are some (and sometimes many) languages which have symmetric coding, where either both constructions are equally coded or both are left uncoded. For the simple case of singular and plural, these two cases can be illustrated by (6a–b).[3]

(6) (a) *Modern Greek (symmetric overt)*

SINGULAR	*vivlí-o*	‘book’
PLURAL	*vivlí-a*	‘books’

(b) *Mandarin Chinese (symmetric zero)*

SINGULAR	*shū*	‘book’
PLURAL	*shū*	‘books’

In languages with symmetric coding, the additional efficiency that is provided by asymmetric coding is not exploited, and a competing constraint takes precedence: In the case of Modern Greek, it is the general preference to express grammatical meanings explicitly, and in Mandarin Chinese, it is the general preference to save coding energy and to leave inferrable meanings unexpressed. We can thus say that English-type languages have an EFFICIENT coding system at the price of asymmetric or non-uniform coding, that Greek has a non-efficient but UNIFORMLY EXPLICIT coding system, and Mandarin Chinese has a non-efficient but UNIFORMLY PARSIMONIOUS system. All three language systems are optimal in their own way, and the form–frequency prediction is relevant only to cases where the coding is asymmetric.[4]

[3] Of course, one would not normally say that Mandarin Chinese has a singular–plural distinction that happens to be unexpressed; instead, one would say that there is no such contrast to begin with. But as noted in the preceding paragraph, we are not talking about language-particular analyses here, but about comparison between languages. From a comparative perspective, Mandarin Chinese has neither overt singular coding nor overt plural coding.

[4] It would be possible to set up Optimality-Theoretic constraint tableaux, where Parsimony would be the highest-ranked constraint in Mandarin, Explicitness (or Faithfulness) would be the highest-ranked constraint in Greek, and Efficiency would be the highest-ranked constraint in English. This

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

What the implicational universal in (4a) predicts not to occur is an asymmetric counter-efficient pattern, where the singular has an overt marker but the plural is left uncoded, as in the Pseudo-Greek pattern in (7), where the plural is uncoded.

(7) *(Pseudo-Greek, hypothetical)*

SINGULAR	*vivlí-**o***	'book'
PLURAL	*vivlí*	'books'

Languages with such a pattern are indeed (virtually) unattested, and similarly for all the patterns seen in Table 1 and in Section 3–7 below.

Thus, such gaps in attested patterns are our explananda, and I claim here that form–frequency correspondences and predictability provide the ingredients for an explanation. The hypothetical diachronic mechanism will be elucidated further in Section 8 below, but the bulk of the paper will consist of listing and briefly discussing a substantial range of universal coding asymmetries.

As I noted earlier, I do not have the space here to provide documentation for each of the asymmetric patterns, and in some cases, more thorough documentation is a task for future research. There is no space here either to justify the claim that one of the construction types occurs more frequently in all languages, as this would go far beyond the scope of an article whose goal is to provide an overview of the general approach. But most of the claims concerning frequency of use are very easy to test, and should thus not be particularly controversial.[5]

Before I get to the various universal patterns, let us briefly consider why the functional-adaptive explanation is proposed at the level of language universals, not at the level of particular languages. Shouldn't the explanation work similarly for language-particular asymmetries? For example, English has two future-tense constructions, with *will* and with *going to* (or *gonna*). These are slightly different semantically, and they are formally asymmetric (*will* is shorter than *gonna*, three vs. four segments). Is it claimed that this is because the *will*-future is more frequent than the *gonna*-future? The answer is no. I make no predictions about such cases, because functional-adaptive explanations only work at the population level (i.e. they explain tendencies found in populations of languages). Language histories are subject to a large number of contingencies, and the adaptive forces are relatively weak. The *gonna*-future could become popular very quickly for social reasons and thus more frequent than the *will*-future (this may well have happened in some

would provide a visualization, but it would also be confusing, because I am not assuming that the explanation resides in constraints of the standard Optimality-theoretic type (= part of the inventory of the innate grammar blueprint). Rather, I am proposing functional-adaptive constraints that operate on language use and influence grammatical systems in a process of cultural evolution (language change, Section 8).

[5] A reviewer asks why the present paper does not contain quantitative evidence to support my claims. The answer is that I concentrate on the more difficult aspects of the overall enterprise here: presenting the conceptual framework (that is often misunderstood) and bringing together a large number of phenomena that are usually discussed in isolation from each other. Interested readers are invited to test the numerous claims that are made in this paper.

varieties of English). But it would not be an exception to a general trend, because there is no general trend for languages to have two distinct future tenses of this sort. It is only when one considers a large enough number of languages that a general trend can emerge, and for a broadly comparative study, one must generally ignore subtle meaning differences such as the difference between the two English future forms. Language-particular patterns are often unrelated to any general trend, and they often arise due to historical contingencies.[6]

In the next two sections (Sections 3 and 4), I give 13 examples of simple meaning pairs that are universally asymmetric. Then in Section 5, I explain the notion of differential coding, and in Sections 6 and 7, I discuss 12 differential-coding pairs.

3. Simple meaning pairs in the nominal domain

3.1. *Singular vs. plural vs. dual*

The most straightforward example is the singular vs. plural opposition. Greenberg (1963) observed that plurals tend to be overtly coded and singulars zero-coded (see Universal 35 in (4a) above), and Greenberg (1966: 32) showed that the singular tends to be more frequent than the plural, while the dual is the least frequent form. The asymmetric coding can be illustrated by Hebrew and Khanty (a Uralic language):

	Hebrew	Khanty
SINGULAR	*yom*	*xot*
PLURAL	*yam-**im***	*xot-**ət***
DUAL	*yom-**ayim***	*xot-**ŋən***
	‘day(s)’	‘house(s)’

(Greenberg also included the dual, thus effectively dealing with a triple rather than with two meanings in opposition; alternatively, the dual could be contrasted with the plural, but the predictions are not different.)

3.2. *Nominative vs. accusative*

Many languages make a distinction between nominative (for the S/A-argument) and accusative (for the P-argument), and Greenberg (1963) noted that the nominative is always at least as short as the accusative (see Universal 38 in (4b) above). The nominative is also more frequent than the accusative, almost by definition, because it occurs in both intransitive and transitive clauses, whereas the accusative only occurs in the latter, e.g.

[6] This is not unlike the situation in other disciplines such as medicine: A particular treatment may be known to be very effective for any population of patients, but this does not mean that any particular patient can be predicted to be cured. There may always be contingent factors that lead to a different outcome in specific cases.

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

	English	German	Quechua
NOMINATIVE	*he*	*Herr Kim*	*wasi* ‘house’
ACCUSATIVE	*hi-**m***	*Herr-**n** Kim*	*wasi-**ta***

3.3. *Allative vs. ablative*

Within the oblique case-markers and adpositions, we find that allative and ablative are asymmetric, with allatives showing a much greater tendency to be zero than ablatives (Stolz, Lestrade & Stolz 2014), and if both are overtly marked, the ablative tends to have a longer shape. Michaelis & APiCS Consortium (2013) provide data from creole and pidgin languages.

	English	Sri Lanka Portuguese	Japanese
ALLATIVE	*to Rome*	*maaket* ‘to the market’	*Tookyoo e* ‘to Tokyo’
ABLATIVE	***from** Rome*	*kaaza **impa*** ‘from home’	*Tookyoo **kara*** ‘from Tokyo’

3.4. *Instrumental vs. comitative*

As can be seen in the many examples in Stolz, Stroh & Urdze (2006), comitative markers are generally longer than instrumental markers when there is a contrast in a language (see also Maurer & APiCS Consortium (2013) for pidgin and creole languages). The comitative is also generally less frequent.

	Welsh	Russian	Hungarian
INSTRUMENTAL	*a* ‘with’	*myš’-ju* ‘with a mouse’	*tol-lal* ‘with a pen’
COMITATIVE	***gyda*** ‘with’	***s** myš’-ju* ‘with a mouse’	*gyerek-**estül*** ‘with a child’

3.5. *Male vs. female occupational terms*

At least in all the languages that I am aware of, female occupational and status terms tend to be longer than male occupational and status terms. Many languages do not have such gender contrasts, but when they make them, it seems that the male form is never longer than the female form. Because special occupational roles (and less often also special status roles) have long been confined to men, it is clear that the male terms have traditionally been more frequent.[7]

	Latin	German	Hungarian	
MALE	*rex*	*König*	*király*	‘king’
FEMALE	*reg-**ina***	*König-**in***	*király-**nő***	‘queen’

[7] Greenberg (1966: 39–40) briefly considers masculine and feminine in adjectives and finds that masculine adjectives tend to be shorter and more frequent.

3.6. *Positive vs. comparative vs. superlative*

In languages that have special comparative and superlative forms of adjectives, the comparative is always derived by a special marker from the positive, and the superlative is often additionally derived from the comparative (Greenberg 1966: 40; Bobaljik 2012). It seems that there is no language with a superlative marker that is shorter than the comparative marker. Greenberg (1966: 41) provides some frequency data showing that the comparative is less frequent than the positive, and the superlative is even less frequent. (The Hungarian and French examples have the same meaning as the English ones.)

	English	Hungarian	French
POSITIVE	*small*	*kis*	*petit*
COMPARATIVE	*small*-***er***	*kis*-***ebb***	***plus*** *petit*
SUPERLATIVE	*small*-***est***	***leg***-*kis*-***ebb***	***le*** *plus* ***petit***

3.7. *Cardinal numerals vs. ordinal numerals*

As has been extensively documented by Stolz (2001: 519), ordinal numerals are normally derived from cardinal numerals, i.e. they consist of an additional marker. Greenberg (1966: 42–43) cites some frequency data showing that they also tend to be much less frequent.

	English	Japanese	Lezgian	
CARDINAL	*seven*	*nanatsu*	*irid*	'7'
ORDINAL	*seven*-***th***	*nanatsu*-***me***	*irid* ***lahaj***	'7th'

4. Simple meaning pairs in the predicational domain

4.1. *Present tense vs. future tense*

Present-tense forms tend to be much more frequent than future-tense forms (Greenberg 1966: 48), and they also tend to be shorter (very often zero-marked, see Bybee 1994).

	English	Latin	Kiribati
PRESENT	*they praise*	*lauda-nt* 'they praise'	*e taetae* 'he speaks'
FUTURE	*they* ***will*** *praise*	*lauda*-***b***-*unt* 'they will praise'	*e* ***na*** *taetae* 'he will speak'

4.2. *Present tense vs. past tense*

Past-tense forms show a similar asymmetry with respect to present-tense forms, both in terms of frequency (Greenberg 1966: 48) and in terms of coding, though less clearly so than future-tense forms.

	Greek	German	Lezgian	
PRESENT	*ksér-is*	*weiß-t*	*či-zwa*	‘know’
PAST	***í**-kser-es*	*wuss-**te**-st*	*či-zwa-**j***	‘knew’

4.3. *Active vs. passive*

Passive verb forms are generally derived from active verb forms, i.e. they contain an additional marker (Haspelmath 1990).

	Italian	Turkish	Russian	
ACTIVE	*vede*	*gör-üyor*	*vidit*	‘sees’
PASSIVE	***è** visto*	*gör-**ül**-üyor*	*vidit-**sja***	‘is seen’

And even in those few cases where active and passive verbs forms both contain markers, as in Latin (e.g. *lauda-s* ‘you (SG) praise’, *lauda-ris* ‘you (SG) are praised’; *lauda-tis* ‘you (PL) praise’, *lauda-mini* ‘you (PL) are praised’), the passive forms tend to be longer. Passive forms are also generally much less frequent than active forms (Greenberg 1966: 46).

4.4. *Basic vs. applicative*

Many languages have applicative verb forms, i.e. forms which contain an additional marker whose function is to indicate that an argument other than the basic P-argument becomes the derived P-argument (Polinsky 2005, Peterson 2007).

	German	Swahili	Dyirbal
BASIC	*fahren* ‘drive’	*andika* ‘write’	*balga-* ‘beat’
APPLICATIVE	***be**-fahren* ‘drive on’	*andik-**ia*** ‘write to’	*balga-**ma**-* ‘beat with’

Somewhat like in the case of the passive, it is part of the definition of ‘applicative’ that it is marked by a special affix, but the derived P-argument is not the most frequent type of argument of the basic verbal meaning, so extra coding on the verb is needed here.

4.5. *Affirmative vs. negative*

Negation is virtually always signaled by an overt negator (Dryer 2005, Miestamo 2005), which might seem trivial, but there are also some languages where there is a special affirmative marker, and there are quite a few languages where tense markers cumulate with affirmative and negative meaning (e.g. Coptic, where the negative past marker is longer than the affirmative past marker). Thus, the generally greater length of negative markers is best explained by the greater rarity of negative clauses.

	Hebrew	English	Coptic
AFFIRMATIVE	*katavti* ‘I wrote’	*I wrote*	*a-f-sôtm* ‘he heard’
NEGATIVE	***lo** katavti* ‘I didn’t write’	*I did**n’t** write*	***mpe**-f-sôtm* ‘he didn’t hear’

4.6. *Disjoint anaphoric vs. reflexive*

Reflexive pronouns are generally longer than anaphoric pronouns which are used when the referent is disjoint from the subject. Again, this can be straightforwardly linked to the much lower frequency of coreference between the object and the subject (Haspelmath 2008c).

	English	Hebrew	M. Chinese	Japanese
DISJOINT	*her*	*oto*	*tā*	Ø
REFLEXIVE	***herself***	*et ʕ**acmo***	(*tā*) ***zìjǐ***	***zibun***

5. Differential coding

So far, we have seen what I have called SIMPLE MEANING PAIRS, where the construction types in opposition are two contrasting grammatical meanings for which the wider grammatical context or lexical subclasses are not immediately relevant.

But form–frequency correspondences are also widely found in situations where a single grammatical meaning is coded differently in different grammatical contexts or in different lexical subclasses. I refer to such situations, where the two constructions in opposition express the same meaning, as DIFFERENTIAL CODING. The best-known example of differential coding is differential object marking (or more precisely differential P flagging, because the core generalization concerns P-arguments of monotransitive verbs, and their flagging by case-markers or adpositions), as illustrated by the well-known contrasts in Spanish and Hebrew that we see in (8) and (9).

(8) (a) Veo la casa.
I.see the house
'I see the house.'
(b) Veo **a** la mujer.
I.see ACC the woman
'I see the woman.'

(9) (a) Kani-ti sefer.
bought-1SG book
'I bought a book.'
(b) Kani-ti **et** ha-sefer.
bought-1SG ACC the-book
'I bought the book.'

Here the patient meaning is expressed in two distinct ways, depending on whether the P nominal is inanimate or animate (in Spanish), or indefinite or definite (in Hebrew). The frequency claim for Spanish concerns the relative frequency of the patient meaning within the sets of all inanimate nominals and all animate nominals: The claim that inanimate Ps are more frequent than animate Ps means that a greater proportion of inanimate nominals have the P role than of animate nouns. The analogous situation holds for Hebrew indefinite and definite Ps.

Another example of differential coding comes from the expression of agent–patient coreference. In a variety of languages, grooming (or INTROVERTED) verbs like ‘wash’ or ‘shave’ show different behaviour from other-directed (or EXTROVERTED) verbs like ‘kill’ or ‘hate’, as seen in English (in 10) and Russian (in 11) (Haiman 1983: 803; Geniušienė 1987; König & Vezzosi 2004).

(10) (a) They shaved. (= ‘They shaved themselves.’)
(b) They hate **themselves**.

(11) (a) Oni myli-s’.
they washed-REFL
‘They washed.’
(b) Oni ubili **sebja**.
they killed themselves
‘They killed themselves.’

In these examples, reflexive marking is differential in that it depends on the lexical subclass (introverted vs. extroverted). The shorter forms (zero-coded in English, reflexive *-s’* in Russian) are used with introverted verbs, because these are more frequently used with agent–patient coreference than extroverted verbs.

One can distinguish two subtypes of differential coding: SPLIT CODING is differential coding that depends on the grammatical context (like differential object marking in Hebrew), while SUBCLASS-CONDITIONED CODING is differential coding that depends on the lexical subclass (like short reflexive coding of introverted verbs in English and Russian).

The explanation of the coding asymmetry is the same as with the simple meaning pairs: Since more inanimate nominals have P-function than animate nominals, hearers are less surprised when they encounter an inanimate P-argument and therefore have less need for special coding.[8] Likewise, since introverted verbs are more often used with agent–patient coreference, hearers are less in need of special reflexive marking than with extroverted verbs.

More generally, in a differential-coding pair, we are dealing with a USUAL ASSOCIATION of a grammatical meaning with a grammatical context or a lexical subclass. The claim is that such usual associations need less coding than unusual associations.

6. DIFFERENTIAL-CODING PAIRS IN THE NOMINAL DOMAIN

In this section, we consider six differential-coding pairs in the nominal domain, which will be followed by six pairs from the predicational domain in Section 7.

[8] A reviewer notes that when I say ‘encounter’ here, I do not mean overt occurrence, because what counts is the frequency and the predictability of the grammatical meanings, not of the overt forms.

6.1. *Accusative marking on inanimate vs. animate P-arguments*

In many languages, P-arguments are coded differentially when they are animate, and as far as I know, the animate P always has a marker, whereas the inanimate P is uncoded (Bossong 1985, 1991; Aissen 2003; Iemmolo 2013).

	Spanish	Armenian
INANIMATE	Ø *la casa* 'house'	*mek* 'another one (inanimate)'
ANIMATE	***a** la mujer* 'woman'	*mek-**i*** 'another one (animate)'

As is well-known, a very similar cross-linguistic pattern is found with definite vs. indefinite P-arguments, e.g. in Hebrew and Turkish. Some languages, such as Hindi-Urdu, have special accusative marking of P-arguments when they are animate or definite. The explanation is the usual association of the P role with inanimate and indefinite arguments (as was clearly stated already by Comrie 1989: 128; Bossong 1991: Sections 2).

6.2. *Ergative marking on 1st/2nd person pronouns vs. full nominals*

In some languages, A-arguments are coded differentially when they are 1st or 2nd person (locuphoric), and this generally means that they lack an overt ergative marker, in contrast to other kinds of arguments, especially full nominals (Dixon 1994: 86).

	Dyirbal	Georgian
1ST PERSON PRONOUN	*ŋadya-Ø*	*me-Ø*
FULL NOMINAL	*yarra-**ŋgu*** 'man'	*mama-**m*** 'father'

This generalization about differential A marking is somewhat less robust than the differential object marking universal, but both belong to a much larger class of role–reference association universals (Haspelmath 2021), which all have the same explanation in terms of form–frequency correspondences, or more specifically in terms of the usual association of high-ranking roles (agent and recipient) with referentially prominent arguments (1st/2nd person, animate, definite, topical).

6.3. *Locative marking on place names vs. inanimate nouns vs. animate nouns*

In a substantial number of languages, locative flagging is differential, such that place names tend to have the shortest coding (often zero, as in Tswana), and animate nouns tend to have the longest coding, with inanimate nouns intermediate between the two (Aristar 1997, Creissels & Mounole 2011, Haspelmath 2019b). In Basque, we see a three-way contrast, while Tswana exemplifies a contrast between place names and inanimates, and Tamil shows a contrast between inanimates and animates.

	Basque	Tswana	Tamil
PLACE NAME	*Bilbo-n* 'in Bilbao'	*Gaborone* 'at Gaborone'	
INANIMATE	*mendi-**tan*** 'at the mountain'	*toporo-**ng*** 'in town'	*N-il*
ANIMATE	*neska-**rengan*** 'at the girl's'		*N-**iṭam***

Again, this is because place names are usually associated with locative use, while this is less common for inanimate nouns, let alone animate nouns. The parallels with differential object marking are so striking that Haspelmath (2019b) calls this pattern 'differential place marking'.

6.4. *Plurative and singulative marking on individualist vs. gregarious nouns*

In some languages, there is a subclass-conditioned difference in singular and plural coding, such that 'individualist nouns' (those that tend to occur with uniplex meaning) have overt plural (plurative) marking, while 'gregarious' nouns (those that are usually associated with multiplex meaning) have singulative (overt singular) marking (*-en* in Welsh, *-ina* in Russian). For detailed discussion, see Haspelmath & Karjus (2017) and Grimm (2018).

		Welsh	Russian	English
INDIVIDUALIST	MULTIPLEX	*cath-od* 'cats'	*koty* 'cats'	*cat-s*
	UNIPLEX	*cath-Ø* 'cat'	*kot-Ø* 'cat'	*cat-Ø*
GREGARIOUS	MULTIPLEX	*moron* 'carrots'	*kartofel'* potatoes'	*salt*
	UNIPLEX	*moron-**en*** 'carrot'	*kartofel-**ina*** 'potato'	***grain of*** *salt*

The suffix *-en* in Welsh is generally regarded as an inflectional singulative marker, while *-ina* in Russian is treated as derivational. In the present context, this difference plays no role, and even the English unit noun *grain* can be seen as playing the same role.

6.5. *Adpossessive marking with inalienable vs. alienable nouns*

In languages where inalienable (kinship and body-part term) nouns behave differently from alienable nouns in adpossessive constructions, the possessive markers are shorter (and typically zero) for the inalienable subclass, because of the usual association of the adpossessive meaning with kinship and body-part nouns (Haspelmath 2017).

	Maltese	Jeli
INALIEN	*id-Ø-i* 'my hand'	*Soma Ø buloni* 'Soma's arms'
ALIEN	*il-ktieb **tiegħ**-i* 'my book'	*Soma **ra** monbilo* 'Soma's car'

6.6. *Definiteness marking with vs. without possessor*

Some languages have differential definiteness marking, omitting the definite article in the presence of an adpossessive nominal.

	German	Hebrew	Welsh
POSSESSED	*mein Ø Buch*	*Ø-sifr-i* 'my book'	*Ø car y meddyg* 'the doctor's car'
UNPOSSESSED	***das*** *Buch*	***ha***-*sefer* 'the book'	***y*** *car* 'the car'

This can be attributed to a form–frequency correspondence as well, as possessed nouns are usually associated with definiteness so that the definite meaning is relatively predictable and the extra coding is not required (Haspelmath 1999).

7. Differential-coding pairs in the predicational domain

7.1. *Reflexive marking on introverted vs. extroverted verbs*

As already discussed in Section 5, some languages have subclass-conditioned differential marking of reflexivity, with introverted verbs (those that are usually associated with coreferential objects) showing shorter coding than extroverted verbs (those that tend to have noncoreferential anaphoric objects) (Haiman 1983: Section 1.2.2; Haspelmath 2008c). This is illustrated by the following examples, which all have the same meaning as the English ones.

	Russian	Dutch	Greek	English
GROOMING	*moet-sja*	*wast zich*	*plen-ete*	*he washes Ø*
EXTROVERTED	*vidit* ***sebja***	*ziet* ***zichzelf***	*vlép-i* ***ton eavtó tu***	*he sees* ***himself***

7.2. *Addressee and 3rd person marking in imperatives vs. indicatives*

There is a usual association of person with mood, in that imperatives tend to have second person subjects and indicatives tend to have 3rd person subjects. Languages often have split coding of bound 2nd and 3rd person person forms, and as predicted, there is a tendency for 2nd person form to be short or zero in imperatives (Aikhenvald 2010: 46), and 3rd person forms to be short or zero (Siewierska 2010).

		Latin	Turkish
IMPERATIVE	2ND	*lauda-Ø* ‘(you) praise!’	*bak-Ø* ‘(you) look’
	3RD	*lauda-**to*** ‘let her praise!’	*bak-**sın*** ‘let her look’
INDICATIVE	2ND	*lauda-v-**isti*** ‘you praised’	*bak-ıyor-**un*** ‘you are looking’
	3RD	*lauda-v-it* ‘she praised’	*bak-ıyor-Ø* ‘she is looking’

7.3. *Attributive and predicative marking on property roots (adjectives) vs. action roots (verbs)*

Property concept roots tend to occur in attributive discourse function, while action roots are usually associated with predicative function. Thus, many languages have a split in how they treat content roots: Property concept roots tend to have short or zero coding in attributive function, while action roots tend to need overt attributive markers (participial affixes or other relativizers). By contrast, in predicative function, property concept roots tend to need special marking by a copula, while action roots do not (Croft 1991: 67).

		German	French	M. Chinese	
ATTR	PROPERTY	*klein-Ø-es Kind*	*petit Ø enfant*	*xiǎo Ø háizi*	‘small child’
	ACTION	*spiel-**end**-es Kind*	*enfant **qui** joue*	*wán **de** háizi*	‘child who plays’
PRED	PROPERTY	*das Kind **is**-t klein*	*l’enfant **est** petit*	*háizi Ø xiǎo*	‘the child is small’
	ACTION	*das Kind spiel-Ø-t*	*l’enfant joue-Ø*	*háizi Ø wán*	‘the child plays’

As Croft (2000) emphasized, it is these kinds of coding splits between different semantic types of roots that are the basis for word-class categorizations, and since they are universal tendencies, the word classes that correspond to the three main discourse functions can be said to be universal (nouns, verbs, adjectives).

7.4. *Causative vs. anticausative marking with automatic vs. costly events*

In some languages, there is a subclass-conditioned difference in noncausal and causal event coding, such that ‘automatic’ verb roots (those that tend to occur with noncausal meaning) have causative marking, while ‘costly’ verb roots (those that are usually associated with causal meaning) have anticausative marking. The basic cross-linguistic pattern is thus completely parallel to that of plurative and singulative marking (Section 6.4).

			French	Russian	Swahili
AUTOMATIC	NONCAUSAL	‘boil (INTR)’	*bouillir*	*kipet’*	*cham-**k**-a*
	CAUSAL	‘boil (TR)’	***faire** bouillir*	*kipj**at**it’*	*cham-**sh**-a*
COSTLY	NONCAUSAL	‘break (INTR)’	***se** casser*	*lomat’-**sja***	*vunj-**ik**-a*
	CAUSAL	‘break (TR)’	*casser*	*lomat’*	*vunj-a*

The semantic classes ‘automatic’ and ‘costly’ are similarly ad hoc as the classes of ‘individualist’ and ‘gregarious’ nouns that are used in Section 6.4 for the number-marking pattern. This is because the cross-linguistic patterns are much less regular than in other cases, and the splits are rather different in different languages. Still, when one looks at a wide range of languages, one sees clear cross-linguistic patterns which were first described in Haspelmath (1993) and linked to universal frequency patterns in Haspelmath et al. (2014).[9]

7.5. *Subordinate clause marking with same vs. different subjects in ‘want’ complements*

‘Want’ complement clauses tend to have the same subject as the main clause, so different-subject complements (‘you want me to eat’) are rare and unexpected. In

[9] An even bigger picture that also explains the distribution of causatives and anticausatives of different types (in particular, analytic vs. synthetic causatives) in terms of form–frequency correspondences is presented in Haspelmath (2016).

accordance with the form–frequency correspondence universal, languages tend to have special subordinate markers for different-subject complements, e.g. *dass* in German and *-te* in Japanese, which are not required in the same-subject situation (see Haspelmath 2013).

	German	Japanese	
SAME SUBJECT	*du willst ess-**en***	*tabe-Ø-tai*	‘you want to eat’
DIFFERENT SUBJECT	*du willst, **dass** ich esse*	*tabe-**te** hosii*	‘you want me to eat’

7.6. *Purpose-clause marking with motion vs. non-motion verbs*

As noted by Schmidtke-Bode (2009: 94), purpose clauses are particularly common (i.e. usually associated) with motion verbs in the main clause, so it is predicted that the purpose-clause marking is generally shorter when the purpose clause modifies a motion-verb clause. Schmidtke-Bode finds that this is generally confirmed, and German provides an illustrative example (zero marking, or pure infinitival marking, is possible only when the main-clause verb is a motion verb).

	German	
MOTION	*sie geht Ø schwimmen*	‘she goes to swim’
NON-MOTION	*sie arbeitet, **um zu** überleben*	‘she works to survive’

8. Paths of change

As I noted in Section 1, the functional-adaptive explanation of the universal tendencies that I propose here relies on the flexibility or malleability of language systems. Our languages are not rigid unchanging systems of rules that we have to simply obey, but they always have some ‘leaks’ or variable usage patterns, and they provide ways of saying things in a novel way. So while the precise pathways along which they change and adapt are usually hard to elucidate, it can at least be imagined how this happens: Language change can be seen as (at least partly) driven by the speakers’ preference for user-friendly utterances, and thus ultimately user-friendly (or adaptive) structures (Keller 1994).

The way in which this works is probably most transparent in lexical change. As I noted briefly in the first section, lexical items show the same strong tendency to exhibit form–frequency correspondences (see the universal in (1)): Frequent words are short, and rare words are long in all languages (e.g. Bentz & Ferrer-i-Cancho 2016). One mechanism by which words that get more frequent are shortened is by clipping (as originally noted by Zipf 1935). For example, advanced intellectual concepts with fairly long names such as ‘mathematics’ and ‘religion’ can become frequent everyday words when they are school subjects, so in colloquial German, the abbreviated forms *Mathe* and *Reli* for these subjects are very common. Clipping is not a regular diachronic process, and clipped forms of rarely used words are very strongly deviant (e.g. **Topo* for *Topologie* ‘topology’, or **Theo* for *Theologie* ‘theology’). Still, when a concept needs to be expressed frequently, it becomes very useful to have a short form for it, and even forms that sound ill-formed at the beginning have a good chance of spreading across the population, regardless of normative perceptions. Thus, usefulness can overcome convention, and new, more adapted conventions can spread across the community.

Another relevant mechanism of change that has been highlighted particularly by Joan Bybee is phonetic reduction. Some words are short because of reduction and overlapping of articulatory gestures, e.g. *gonna* (from *going to*), *don't* (from *do not*) (e.g. Bybee 2015: Section 6.6). Phonetic reduction plays a role in explaining coding asymmetries in some cases, e.g.

(12)	(a)	English	*mine*	(independent possessive pronoun)
			my	(bound possessive pronoun)
	(b)	Polish	*śpiewa-sz*	[sing-2SG] ‘you sing’
			śpiewa-Ø	[sing-3SG] ‘she sings’

In both pairs, the second member is more frequent and shorter and derives from a former longer form (*my* < *mine*; *śpiewa* < **śpiewa-t*), apparently by phonetic reduction.

However, neither clipping nor phonetic reduction are the mechanisms of change in the great majority of asymmetric patterns that we saw in Sections 3–7 and Sections 6–7. In most cases, the asymmetries are the result of differential development of a new construction. For example, in the pairs in (13), the first member of the pair has a novel element that was created at some point, while the second member of the pair did not develop a marker for the relevant meaning.

(13)	(a)	English	*they will praise*	(future; *will* from ‘want’)
			they Ø praise	(present tense)
	(b)	French	*se casser*	(intransitive ‘break’; *se* from ‘(it)self’)
			Ø casser	(transitive ‘break’)
	(c)	Russian	*vidit-sja*	(passive ‘is seen’; *-sja* from ‘(it)self’)
			vidit-Ø	(active ‘sees’)

In many other cases, we do not know how the marker developed because it is very old – e.g. the comparative marker in *small-er* (Section 3.6) or the copula in *the child is small* (Section 7.3), or the 3rd person imperative marker in Turkish *bak-sın* ‘let her look’ (Section 7.2). But it is quite likely that these asymmetric patterns, too, arose by differential development.

Similar is the situation of differential expansion of a new marking pattern (Haspelmath 2008b: 207). The newly developed definite article did not spread to possessed nominals in German (14a), and the newly developed possessive marker *tiegħ-* did not spread to inalienable nouns in Maltese (14b).

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

(14)	(a)	German	*das Buch*	‘the book’	
			mein Ø Buch	‘my book’	
				(no definite article, recall Section 6.6 above)	
	(b)	Maltese	*il-ktieb tiegħ-i*	[the-book of-1SG]	‘my book’
			id-Ø-i	[hand-Ø-1SG]	‘my hand’
				(recall Section 6.5 above)	

Another possible way in which asymmetric coding may come about is by eliminating a coding contrast only where it is not particularly needed. In Middle High German, all masculine nouns of the *-n* class made a nominative–accusative distinction, regardless of animacy. In Modern German, this distinction was given up for inanimates, but not for animates such as *Affe* ‘ape’, so that we get a situation with differential accusative marking, as seen in (15) (see Nübling 2008: 303–306).

(15)		NOMINATIVE	ACCUSATIVE	
	Middle High German	*knote-Ø*	*knote-n*	‘knot’
		affe-Ø	*affe-n*	‘ape’
	Modern German	*Knoten*	*Knoten*	‘knot’
		Affe-Ø	*Affe-n*	‘ape’

There are thus multiple ways in which asymmetric coding can come about in a language. This shows that the asymmetric coding patterns are not due to tendencies inherent in the kinds of changes that lead to them, but to factors favouring particular results (see Section 9.4 for more discussion).

9. Frequency and predictability in the causal chain

In this section, I discuss a few more aspects of the efficiency-based explanation of asymmetric coding tendencies.

9.1. *Markedness and iconicity*

In earlier work, I examined two alternative explanations for asymmetric coding and argued that they were less comprehensive or did not provide the necessary causal links.

In Haspelmath (2006), I argued that markedness does not provide an explanation, because there is no unitary and generally recognized markedness concept, and if we adopted a single concept of markedness as a representational feature of universal grammar, we would not have a causal link. An explanation that provides causal links is preferable to an explanation that replaces the explanandum by another unexplained concept elsewhere for which there is no independent evidence.

In Haspelmath (2008a), I argued that iconicity (or the tendency for meaning–form matching) cannot generally work as an explanation of the universals of asymmetric coding, because in many cases there are no meaning differences but we still find coding asymmetries. This is the case in all differential-coding pairs (Sections 5–7), where the same grammatical meaning is expressed in both construction types. For asymmetries like accusative marking of animate or inanimate objects, locative marking on place names and ordinary nouns, and attributive vs. predicative use of property words (adjectives), frequency-based predictability makes the right prediction, but iconicity does not seem to make any prediction. And in the case of positive vs. comparative (Section 3.6), it has often been argued that the comparative is actually simpler semantically.

Another possible alternative explanation might reverse the causal chain in (3) (frequency → predictability → shortness) and claim that higher frequency of use of shorter forms is due to the shortness of the forms. For example, the singular might be more frequent than the plural because its forms are shorter. This explanation is not only unintuitive (especially for coding asymmetries like affirmative–negative or positive–comparative), but it makes the wrong prediction that when a language has symmetric coding (e.g. when it has both a present-tense marker and a future-tense marker, both equally long), there should be no frequency difference. But in fact, the frequency differences are independent of the coding, as they are found in all languages, regardless of their coding patterns.

9.2. *Coding efficiency and predictability*

The explanation that I propose relies on the recognition that language systems tend to be efficient, i.e. they help the users make good use of scarce resources (speaker articulations) for the desired result of communication (understanding the speaker's intentions). There should thus be a balance between parsimony and clarity: articulations should be reduced only to the extent that this does not threaten comprehension.

Such a balance or trade-off between ease of articulation and ease of decoding has often been postulated in linguistics, and there should be no need to justify the general principle here (see e.g. von der Gabelentz 1891; Langacker 1977; Lindblom & Maddieson 1988; Fedzechkina, Jaeger & Newport 2012: 1; Gordon 2016: 17; Kemp, Xu & Regier 2018).

Now of course a key factor that aids comprehension is predictability: If the content of a message is not surprising, the message can be abbreviated. Speakers can afford to use short shapes or zero coding for predictable meanings, but they have to make a greater coding effort for unpredictable meanings.

If context is kept constant, higher-frequency meanings are more predictable than lower-frequency meanings because of their frequency: It is less surprising if my interlocutor uses a present-tense form than if she uses a future-tense form, just as it is less surprising if she talks about *Mathe*(*matik*) than if she talks about *Topologie* (recall Section 7 above). Thus, as was already noted in Section 1, the causal chain goes from high frequency to predictability, and from predictability to short coding.

The predictability-based explanation is confirmed by considering phenomena that involve predictability but not (necessarily) high frequency: I mean contextually

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

high predictability of meanings.[10] For example, when an utterance is about a referent that has occurred just before, it is of course highly predictable (or accessible, Ariel 1990). In such situations, languages generally use short forms or zero, as is illustrated by the examples in (16).

(16) *Contextual predictability*
    (a) Short form for predictable referent:
        The girl went to the river; **she** looked for fish.
    (b) Zero coding for predictable referent:
        The girl went to the river and **Ø** looked for fish.

In these cases, the counter-efficient pattern, with the short form where the referent is unpredictable, is impossible (**She*$_{\text{i}}$ *went to the river and the girl*$_{\text{i}}$ *looked for fish*). That personal pronouns and other anaphoric forms are generally shorter than full nominals is not in doubt but is rarely highlighted, because linguists have not often considered coding length as an important concept for understanding morphosyntactic patterns. But in the present context, these considerations support the explanation of universal asymmetries in terms of coding efficiency and predictability. The causal chain sketched in (3) can thus be expanded into the chain in Figure 1, showing two distinct causes of predictability, but a uniform result of predictability.

9.3. *What is the ultimate cause?*

When presenting the frequency- and predictability-based explanation of coding asymmetries over the years, I have often heard the comment that frequency cannot be the ultimate cause of the cross-linguistic patterns. Frequency itself must be caused by something, and this is not represented in the causal chain in Figure 1. What is it that causes the higher frequency of use of singular as opposed to plural, of present tense as opposed to future tense, of affirmative as opposed to negative, and so on? What causes inanimate nominals to occur in P function more frequently than animate nominals? Maybe this factor, whatever it is, could independently cause both frequency/predictability and shortness of coding, so that the causal chain would be different – as in Figure 2, and not as in Figure 1.

This is a possibility that must be taken seriously, but it is quite unclear what this factor X might be. Mayerthaler (1981: 136–140) proposed that markedness causes both high frequency and shortness of coding, but ‘markedness reversals’ show that this cannot be the explanation (see Haspelmath 2006: 44), and it is completely unclear how the abstract feature of markedness would lead to greater frequency of

[10] Some psycholinguists have claimed that contextual or local predictability is a better predictor of a variety of effects, including word length (e.g. Piantadosi, Tily & Gibson 2011), than frequency of use. But it seems that the two measures are often highly correlated, and at the level of resolution needed for explaining cross-linguistic tendencies in grammatical coding, the differences do not seem to matter much.

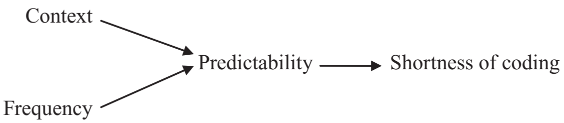

Figure 1
The causal chain leading to shortness of coding.

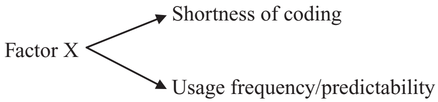

Figure 2
An alternative conception of the causal chain.

use. I do not know of any other candidate for such a factor X, so I will not discuss this possibility further here.

Ultimately, we would of course like to know the causes of frequency asymmetries, but the causes seem to be quite diverse. For example, the greater frequency of present tense vs. future tense seems to have to do with what we know about the present and about the future: The future is much less certain, so we cannot talk much about it. The greater frequency of male profession nouns vs. female profession nouns must have to do with frequency in the world: Male occupations are traditionally more diversified, so male profession nouns are (or were) used more frequently. But the greater frequency of singular vs. plural must have to do with our cognitive preferences – it cannot be due to the world (which does not contain more singular entities than pluralities) or to our knowledge of the world. It seems that we simply prefer to talk about individual entities.

In contrast to the diverse origins of the frequency asymmetries, their results are very uniform, and this is what the present paper aims to explain.

9.4. *Ways of demonstrating the causal link*

The focus of this paper has been on presenting a wide range of phenomena from the world's languages, some of which have been studied extensively and are well known, highlighting that they are all special cases of the overarching form–frequency generalization of (2). I have also discussed a proposed causal link between frequency and asymmetric coding, in line with some earlier work. In this section, I will briefly consider two points of criticism that were raised by two reviewers.

On the one hand, one might ask what the evidence is that the disparate kinds of changes that I surveyed in Section 8 are indeed motivated by a preference for efficient coding. Cristofaro (2019: 27) thinks that 'recurrent grammatical configurations do not appear to arise because of principles that that favor those particular configurations in themselves'. And indeed, historical linguists have often attributed diachronic changes to processes such as analogy, context-driven reinterpretation, metonymization, or grammaticalization, rather than to 'efficientization' processes. One might therefore say that the observed outcomes are not driven by functional adaptation, but by these classical processes and by the properties of the typical sources (Cristofaro 2017 calls this the 'source-oriented' approach).

And indeed, language change often appears to be regular, and one might say that in addition to result-oriented functional-adaptive constraints, there are also MUTATIONAL CONSTRAINTS (constraints on possible changes or possible diachronic sources) that lead to observed universals (Haspelmath 2019a). However, as we saw in Section 8, there are a wide range of disparate paths of change that lead to uniform outcomes. This kind of multi-convergence of diverse changes cannot be explained by a source-oriented approach. Cristofaro (2019: 38) notes that 'different diachronic processes can all lead to the same synchronic outcome for different reasons', but this would be an accidental coincidence, and the universal claims of this paper go far beyond such possible coincidences. It may seem surprising that the resulting cross-linguistic regularities are unrelated to presumed regularities of change, but I would point to the analogous situation biological evolution, where functional-adaptive changes do not arise from regular changes either (but from functional selection; see Nettle 1999).

On the other hand, one might ask whether we can go beyond observing the correlations between asymmetric coding and frequency, and test the causal hypothesis experimentally. This has in fact been done in a number of studies using miniature artificial language learning paradigms. Thus, Fedzechkina et al. (2012) reported that learners restructured an artificial language with differential case-marking in such a way as to match the cross-linguistically common patterns, and Kurumada & Grimm (2019) found that learners preferentially produced plural marking on nouns that are less probable to occur with plural meaning, in line with the cross-linguistic tendencies (recall Section 6.4 above).[11] There are also a number of experimental studies of languages with optional marking that point in the same direction, e.g. Kurumada & Jaeger's (2015) experiments on optional case-marking in Japanese. These works have not been influential among linguists who are not specialized in psycholinguistics, but their results are fully compatible with the claims of this paper. Of course, there are many new issues that arise once one considers the precise psychological mechanisms that might be responsible for the speakers' behaviour (see Jaeger & Buz 2018 for a comprehensive survey). It is hoped that the present paper will contribute to bringing this line of work from psycholinguistics more closely together with comparative grammar research.

It is important to keep in mind, however, that even though not all questions about the causal connections may have been answered, there is no competing theory that

[11] Kanwal et al. (2017) got analogous results when they studied word length in a similar experimental design, 'supporting Zipf's conjecture that the Principle of Least Effort can explain this universal feature of word length distributions'.

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

can lay claim to explaining even a fraction of the phenomena of Section 3–7. In practice, the main alternative perspective that this paper argues against is that all these phenomena should be discussed (and perhaps explained) in isolation from each other.

10. Some history

The ideas presented here are not quite original, so I will briefly relate them to the most important earlier general work on asymmetric coding in grammar: Greenberg (1966), Croft (1990, 2003), and Hawkins (2004).

Greenberg (1966) was the first to note that a wide range of simple meaning pairs (some of which we saw in Section 3–4) are asymmetric in a number of ways. He did not mention asymmetric coding in differential-coding pairs, but instead emphasized the apparent parallels with asymmetries in phonological segment inventories. He saw both asymmetric coding ('zero expression of the unmarked category') and frequency of use as manifestations ('criteria') of markedness, and did not focus on possible explanations of the patterns he discovered (much as in his 1963 paper, where explanations were secondary). However, he briefly noted that while the greater frequency of 'unmarked' phonological segments is the result of their other unmarked properties, 'the role of frequency in grammar-semantics' seems to be primary (Greenberg 1966: 65).

Haiman's (1983) paper (see also Haiman 1985) was important because it discussed a diverse range of phenomena such as adpossessive marking (see Section 6.5 above), reflexive marking (see Section 7.1 above), and causatives (see Section 7.4 above) in terms of 'iconicity' and 'economy' and highlighted the role of tradeoffs ('competing motivations').

Croft (1990: Chapter 4) followed Greenberg closely and was inspired by Haiman, but in the 2003 version of the book (Croft 2003), he no longer treats frequency as simply another criterion of markedness. The new version claims that frequency provides a 'deeper explanation' for the observed economy and iconicity effects (where 'economy' refers to parsimony, and 'iconicity' to explicitness in my terms). He notes that the explanation is the same as Zipfian explanation for the correlation between shortness and frequency (Croft 2003: 112). He also deals with differential-coding pairs at length (Chapter 6), but he discusses these under the confusing heading of 'typological prototypes', and the relation between these patterns and usage frequency does not become very clear. Still, Croft's work is the most important precursor to the present proposal.

Hawkins (2004: Section 3.2) formulates a very general principle ('Minimize Forms') which makes similar predictions to the explanations of asymmetric coding that I have proposed here. However, he only discusses case, number and gender asymmetries, and has almost nothing to say about differential coding (see also the summary in Hawkins 2014: Section 2.2).[12] Outside of the typological community,

[12] He only mentions differential object marking briefly, but his explanation is identical to the one given here ('inanimacy and indefinitenes permit the inference to "objecthood" because of their

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

the idea of a tradeoff between speaker effort and robust information transmission has also been pursued by a number of researchers, e.g. in phonology, variational corpus linguistics, and psycholinguistics (Aylett & Turk 2004, Jaeger & Tily 2011, Hall et al. 2018, Gibson et al. 2019, Levshina 2019).

Thus, the ideas and claims summarized in this paper are not completely novel, and the only claim I make is that this is the right way of explaining a large number of cross-linguistic patterns that are still often discussed in very different terms (apparently because the Greenberg–Croft–Hawkins insights have not become widely known throughout the discipline yet).

11. Conclusion

To summarize, I have claimed that when two minimally different grammatical construction types differ in frequency across languages, they will also show a universal tendency to exhibit asymmetric coding, i.e. the more frequently used construction will tend to be coded by a shorter form (or by zero), while the less frequently used construction will be coded by a longer form.

I have listed 25 pairs of construction types for which this claim either has been substantiated in the earlier literature or seems very plausible. I am planning to document quite a few further cases of coding asymmetries in future work.

I have proposed that the universal coding asymmetries can be explained as caused by the functional-adaptive force of coding efficiency. This means that speakers, and therefore also languages, must make a tradeoff between the conflicting pressures of parsimony and explicitness. While some languages in some patterns show fully explicit coding, and some languages in some patterns show no overt coding (i.e. maximal parsimony), many languages make an efficient compromise, where only the less predictable information gets overt coding, while the more predictable information is left uncoded (or is coded with fewer segments). Counter-efficient patterns are virtually unattested.

The tradeoff that languages make can be seen as a consequence of the tradeoffs that speakers make in language use, because through language change (a kind of cultural evolutionary process), languages can adapt to the needs of the speakers. There are multiple pathways by which efficient coding patterns can arise, which means that we have convergent cultural evolution.

For quite a few of the individual coding types, there is a rich earlier literature, and sometimes rather different explanations have been proposed. There was no space in this article to compare my explanation with other explanations, but some of my earlier papers include some specific comparisons (e.g. Haspelmath 1999: Section 6; Haspelmath 2008a, 2008c: Sections 2 and 8; Haspelmath et al. 2014: Section 3;

frequent association, and this inference permits zero object marking'). Hawkins (2004: Section 3.2.3) is also worth reading because he relates his Minimize Form to the expression of contextually predictable information (accessibility in Ariel's (1990) terms) and to Levinson's (2000) informativeness principle, as well as Relevance Theory. These further connections are worth exploring, but beyond the scope of the present paper.

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

Haspelmath 2017; Haspelmath & Karjus 2017: Section 3). Here I have limited myself to a few remarks on iconicity and markedness explanations, but especially for the domain of argument coding and reflexivization, there is also a voluminous literature from a generative perspective where the attempt is sometimes made to explain implicational universals. I claim that all these explanations can and should be replaced by the functional-adaptive explanation in terms of predictability and coding efficiency that I have advanced here,[13] and I hope that future work will elaborate further on these ideas.

REFERENCES

Aikhenvald, Alexandra. 2010. *Imperatives and commands*. Oxford: Oxford University Press.

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory* 21.3, 435–483.

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.

Aristar, Anthony Rodrigues. 1997. Marking and hierarchy types and the grammaticalization of case-markers. *Studies in Language* 21.2, 313–368.

Aylett, Matthew & Alice Turk. 2004. The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47.1, 31–56.

Bentz, Christian & Ramon Ferrer-i-Cancho. 2016. Zipf's law of abbreviation as a language universal. In Christian Bentz, Gerhard Jäger & Igor Yanovich (eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. Tübingen: University of Tübingen. doi:http://dx.doi.org/10.15496/publikation-10057.

Bobaljik, Jonathan David. 2012. *Universals in comparative morphology: Suppletion, superlatives, and the structure of words*. Cambridge, MA: MIT Press.

Bossong, Georg. 1985. *Differenzielle Objektmarkierung in den neuiranischen Sprachen*. Tübingen: Narr.

Bossong, Georg. 1991. Differential object marking in Romance and beyond. In Douglas Kibbee & Dieter Wanner (eds.), *New analyses in Romance linguistics*, 143–170. Amsterdam: John Benjamins.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, Joan L. 1994. The grammaticization of zero: Asymmetries in tense and aspect systems. In William Pagliuca (ed.), *Perspectives on grammaticalization* (Current Issues in Linguistic Theory 109), 235–254. Amsterdam: John Benjamins.

Bybee, Joan L. 2015. *Language change*. Cambridge: Cambridge University Press.

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

Creissels, Denis & Céline Mounole. 2011. Animacy and spatial cases: Typological tendencies, and the case of Basque. In Seppo Kittilä, Katja Västi & Jussi Ylikoski (eds.), *Case, animacy and semantic roles*, 155–182. Amsterdam: John Benjamins.

Cristofaro, Sonia. 2017. Implicational universals and dependencies. In N. J. Enfield (ed.), *Dependencies in language: On the causal ontology of linguistic systems*, 9–22. Berlin: Language Science Press. http://langsci-press.org/catalog/book/96.

Cristofaro, Sonia. 2019. Taking diachronic evidence seriously: Result-oriented vs. source-oriented explanations of typological universals. In Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis & Ilja A. Seržant (eds.), *Explanation in typology*, 25–46. Berlin: Language Science Press. http://langsci-press.org/catalog/book/220.

[13] Note that by 'explanation', I mean the explanation of universal tendencies, not the characterization of language-particular grammatical regularities (confusingly, the term 'explanation' is sometimes also used for the latter, e.g. Newmeyer 2017). Language-particular analysis is not relevant in the present context.

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

Croft, William. 1990. *Typology and universals*. Cambridge: Cambridge University Press.

Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago, IL: University of Chicago Press.

Croft, William. 2000. Parts of speech as language universals and as language-particular categories. In Petra M. Vogel & Bernard Comrie (eds.), *Approaches to the typology of word classes*, 65–102. Berlin: Mouton de Gruyter.

Croft, William. 2003. *Typology and universals*, 2nd edn. Cambridge: Cambridge University Press.

Diessel, Holger. 2019. *The grammar network*. Cambridge: Cambridge University Press.

Dixon, R. M. W. 1994. *Ergativity*. Cambridge: Cambridge University Press.

Dryer, Matthew S. 2005. Negative morphemes. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 454–457. Oxford: Oxford University Press. http://wals.info/chapter/112.

Fedzechkina, Maryia, T. Florian Jaeger & Elissa L. Newport. 2012. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences* 109.44, 17897–17902. doi:10.1073/pnas.1215776109.

Geniušienė, Emma. 1987. *The typology of reflexives*. Berlin: Mouton de Gruyter.

Gibson, Edward, Richard Futrell, Steven T. Piandadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen & Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences* 23.5, 389–407.

Gordon, Matthew Kelly. 2016. *Phonological typology*. Oxford: Oxford University Press.

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

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

Grimm, Scott. 2018. Grammatical number and the scale of individuation. *Language* 94.3, 527–574.

Haiman, John. 1983. Iconic and economic motivation. *Language* 59.4, 781–819.

Haiman, John. 1985. *Natural syntax: Iconicity and erosion*. Cambridge: Cambridge University Press.

Hall, Kathleen Currie, Elizabeth Hume, T. Florian Jaeger & Andrew Wedel. 2018. The role of predictability in shaping phonological patterns. *Linguistics Vanguard* 4(s2). doi:10.1515/lingvan-2017-0027.

Haspelmath, Martin. 1990. The grammaticization of passive morphology. *Studies in Language* 14.1, 25–72.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In Bernard Comrie & Maria Polinsky (eds.), *Causatives and transitivity*, 87–120. Amsterdam: John Benjamins.

Haspelmath, Martin. 1999. Explaining article–possessor complementarity: Economic motivation in noun phrase syntax. *Language* 75.2, 227–243.

Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42.1, 25–70.

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

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

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

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86.3, 663–687.

Haspelmath, Martin. 2013. On the cross-linguistic distribution of same-subject and different-subject ‘want’ complements: Economic vs. iconic motivation. *SKY Journal of Linguistics* 26, 41–69.

Haspelmath, Martin. 2016. Universals of causative and anticausative verb formation and the spontaneity scale. *Lingua Posnaniensis* 58.2, 33–63.

Haspelmath, Martin. 2017. Explaining alienability contrasts in adpossessive constructions: Predictability vs. iconicity. *Zeitschrift für Sprachwissenschaft* 36.2, 193–231.

Haspelmath, Martin. 2019a. Can cross-linguistic regularities be explained by constraints on change? In Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis & Ilja A. Seržant (eds.), *Explanation in typology*, 1–23. Berlin: Language Science Press. doi:10.5281/zenodo.2583804

Haspelmath, Martin. 2019b. Differential place marking and differential object marking. *STUF–Language Typology and Universals* 72.3, 313–334.

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

Haspelmath, Martin. 2021. Role–reference associations and the explanation of argument coding splits. To appear in *Linguistics*.

Haspelmath, Martin, Andreea Calude, Michael Spagnol, Heiko Narrog & Eli f Bamyacı. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation1. *Journal of Linguistics* 50.3, 587–625.

Haspelmath, Martin & Andres Karjus. 2017. Explaining asymmetries in number marking: Singulatives, pluratives, and usage frequency. *Linguistics* 55.6, 1213–1235.

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

Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. New York: Oxford University Press.

Iemmolo, Giorgio. 2013. Symmetric and asymmetric alternations in direct object encoding. *STUF–Language Typology and Universals* 66.4, 378–403.

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61.1, 23–62.

Jaeger, T. Florian & Esteban Buz. 2018. Signal reduction and linguistic encoding. In Eva M. Fernández & Helen Smith Cairns (eds.), *The handbook of psycholinguistics*, 38–81. Hoboken, NJ: Wiley.

Jaeger, T. Florian & Harry Tily. 2011. On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science* 2.3, 323–335.

Kanwal, Jasmeen, Kenny Smith, Jennifer Culbertson & Simon Kirby. 2017. Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition* 165, 45–52.

Keller, Rudi. 1994. *On language change: The invisible hand in language*. London: Routledge.

Kemp, Charles, Yang Xu & Terry Regier. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics* 4.1, 109–128.

König, Ekkehard & Letizia Vezzosi. 2004. The role of predicate meaning in the development of reflexivity. In Walter Bisang, Nikolaus Himmelmann & Björn Wiemer (eds.), *What makes grammaticalization? A look from its fringes and its components*, 213–244. Berlin: de Gruyter.

Kurumada, Chigusa & Scott Grimm. 2019. Predictability of meaning in grammatical encoding: Optional plural marking. *Cognition* 191, 103953. https://doi.org/10.1016/j.cognition.2019.04.022.

Kurumada, Chigusa & T. Florian Jaeger. 2015. Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83. 152–178.

Langacker, Ronald W. 1977. Syntactic reanalysis. In Charles Li (ed.), *Mechanisms of syntactic change*, 57–139. Austin, TX: University of Texas Press.

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Levshina, Natalia. 2019. *Towards a theory of communicative efficiency in human languages*. Habilitation thesis, Leipzig University.

Lindblom, Björn & Ian Maddieson. 1988. Phonetic universals in consonant systems. In Larry M. Hyman & Charles N. Li (eds.), *Language, speech, and mind*, 62–78. London: Routledge.

Maurer, Philippe & APiCS Consortium. 2013. Comitatives and instrumentals. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *Atlas of pidgin and creole language structures*, 276–280. Oxford: Oxford University Press. https://apics-online.info/parameters/70.

Mayerthaler, Willi. 1981. *Morphologische Natürlichkeit*. Wiesbaden: Athenaion.

Michaelis, Susanne Maria & APiCS Consortium. 2013. Going to named places. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *Atlas of pidgin and creole language structures*, 314–317. Oxford: Oxford University Press. http://apics-online.info/parameters/79.

Miestamo, Matti. 2005. *Standard negation: The negation of declarative verbal main clauses in a typological perspective*. Berlin: Mouton de Gruyter.

Nettle, Daniel. 1999. Functionalism and its difficulties in biology and linguistics. In Mike Darnell, Edith A. Moravcsik, Frederick J. Newmeyer, Michael Noonan & Kathleen Wheatley (eds.), *Functionalism and formalism in linguistics*, vol. I, 445–467. Amsterdam: John Benjamins.

Newmeyer, Frederick J. 2017. Formal and functional explanation. In Ian Roberts (ed.), *The Oxford handbook of universal grammar*. Oxford: Oxford University Press.

Nübling, Damaris. 2008. Was tun mit Flexionsklassen? Deklinationsklassen und ihr Wandel im Deutschen und seinen Dialekten. *Zeitschrift für Dialektologie und Linguistik* 75.3, 282–330.

Peterson, David A. 2007. *Applicative constructions*. Oxford: Oxford University Press.

Piantadosi, Steven T., Harry Tily & Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108.9, 3526–3529.

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press

Polinsky, Maria. 2005. Applicative constructions. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 442–445. Oxford: Oxford University Press. http://wals.info/chapter/109.

Reinhart, Tanya & Eric Reuland. 1993. Reflexivity. *Linguistic Inquiry* 24.4, 657–720.

Schmidtke-Bode, Karsten. 2009. *A typology of purpose clauses*. Amsterdam: John Benjamins.

Siewierska, Anna. 2010. Person asymmetries in zero expression and grammatical function. In Franck Floricic (ed.), *Essais de typologie et de linguistique générale: Mélanges offerts à Denis Creissels*, 471–485. Lyon: ENS Éditions.

Stolz, Thomas. 2001. ORDINALIA–Linguistisches Neuland: Ein Typologenblick auf die Beziehung zwischen Kardinalia und Ordinalia und die Sonderstellung von EINS und ERSTER. In Birgit Igla & Thomas Stolz (eds.), *Was ich noch sagen wollt... A multilingual Festschrift for Norbert Boretzky on occasion of his 65th birthday*, 507–530. Berlin: Akademie.

Stolz, Thomas, Sander Lestrade & Christel Stolz. 2014. *The crosslinguistics of zero-marking of spatial relations*. Berlin: De Gruyter Mouton.

Stolz, Thomas, Cornelia Stroh & Aina Urdze. 2006. *On comitatives and related categories: A typological study with special focus on the languages of Europe*. Berlin: Mouton de Gruyter.

von der Gabelentz, Georg. 1891. Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse. Leipzig: C. H. Tauchnitz. (2016 republication: https://langsci-press.org/catalog/book/97

Zipf, George Kingsley. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.

Author's address: *Max Planck Institute for Evolutionary Anthropology,*
Leipzig University, Deutscher Platz 6, D-04103 Leipzig, Germany
martin_haspelmath@eva.mpg.de

https://doi.org/10.1017/S0022226720000535 Published online by Cambridge University Press