Frequency effects in morphology 12

The ways in which speakers use language have a profound influence on language structure, and frequency is one of the most important sources for system-external explanation of language structure. In fact, we have already seen examples in this book of frequency affecting the content of the lexicon (Section 4.3), productivity (Sections 6.4.1–6.4.2) and word-class shift (Chapter 8). In this chapter we explore how frequency matters for language structure, and why. Frequency influences word structure in many ways, but one of the most striking effects is found in inflection, where frequency asymmetries result in asymmetrical structural behaviour of various kinds. We look at some examples of this interaction.

12.1 Asymmetries in inflectional values

In inflectional systems, we often observe asymmetries in the behaviour of inflectional values that belong to the same inflectional feature, including number (singular versus plural), case (nominative versus accusative), voice (active versus passive) and polarity (affirmative versus negative).

12.1.1 Frequent and rare values

Frequency differences among some common inflectional values are summarized in Table 12.1, where '>' means 'is more frequent than'. It should be noted that not every word in a language will exhibit these frequency asymmetries. These generalizations should instead be taken as describing the overall pattern of a language. Also, not all languages have inflection for all these features, but the claim is that, when a language has inflection for one of these features and values, it will conform to the generalization expressed in the table.

Feature	Values, ordered by frequency
number	singular > plural > dual
case	nominative > accusative > dative
person	3rd > non-3rd (1st/2nd)
degree	positive > comparative > superlative
voice	active > passive
mood	indicative > subjunctive
polarity	affirmative > negative
tense	present > future

Table 12.1 Frequent and rare values

The generalizations in Table 12.1 can be illustrated by examining the results produced by counting usage of number values in six languages:

(12.1)

	Singular	Plural	Dual	Number of nouns
French	74.3%	25.7%		1,000
Latin	85.2%	14.8%		8,342
Russian	77.7%	22.3%		8,194
Sanskrit	70.3%	25.1%	4.6%	93,277
Slovene	72.5%	26.9%	0.6%	11,711
Upper Sorbian	64%	30%	6%	unknown

(Corbett 2000: 281; data partly from Greenberg 1966: 31–2)

The differences between languages that we see here could be due to slight differences in the meanings of the number values, or they might simply be due to the genre or style of the text chosen. Ideally, the token frequency of inflectional values should be counted in a text that is representative of the everyday spoken language in the community, and finding such representative texts is not straightforward. But, fortunately, the asymmetries in the usage of number values are so robust that the same result is generally obtained, no matter which texts we look at. So we can safely say that the singular has a higher frequency than the plural.

Why should such frequency asymmetries exist? To start with, the nominative can be expected to be more frequent than the accusative, at least in languages that do not allow unexpressed arguments, because all verbs require a nominative argument (i.e. a subject), but only transitive verbs also have an accusative. Similarly, the subjunctive must be rarer than the indicative because subjunctives are used primarily in subordinate clauses, and a subordinate clause presupposes a main clause with, at least typically, an indicative verb. However, the ultimate reason for the different frequencies

of different inflectional values is outside language. Some expressions are more frequent simply because humans find them more relevant: we all talk more about singular entities than about plural entities, more about third persons and things than about speech act participants (first/second person), more about present events than about future events, and so on. The linguist has no privileged skills for explaining these preferences, so we will not discuss them further. Instead, we will focus on structural properties that correlate with frequency.

12.1.2 The correlation between frequency and shortness

Quite generally, frequent expressions tend to be short in human languages. Frequent words are shorter than rare words. For example, in French the 10 most frequent words are *de*, *le*, *la*, *et*, *les*, *des*, *est*, *un*, *une*, *du*, and long words like *éléphant* or *questionnaire* are used rarely.

Even more strikingly, frequently used inflectional values may not be expressed overtly at all but are left to be inferred from the context – i.e. they sometimes show zero expression. This is just one more manifestation of the correlation between frequency and shortness. As an example, consider the partial inflectional paradigm of regular nouns in Udmurt, given in (12.2).

(12.2)

	SINGULAR	PLURAL	
NOMINATIVE	*val*	*valjos*	'horse(s)'
ACCUSATIVE	*valez*	*valjosty*	'horse(s) (dir. obj.)'
ABLATIVE	*valleś*	*valjosleś*	'from the horse(s)'
ABESSIVE	*valtek*	*valjostek*	'without the horse(s)'

(Perevoščikov 1962: 86–7)

In this paradigm, the more rarely used cases, ablative and abessive, have a longer form than the more frequently used accusative. The nominative and the singular are the shortest: they are both expressed by zero. This Udmurt paradigm is quite typical of inflectional systems. Zero expression is found in frequent values, and when two contrasting values are both overtly coded, typically the more frequently used value has the shorter expression. Two more examples from verbal inflection are given in (12.3) and (12.4).

(12.3)

Tzutujil	COMPLETIVE	INCOMPLETIVE	POTENTIAL
1SG	*x-in-wari*	*n-in-wari*	*xk-in-wari*
2SG	*x-at-wari*	*n-at-wari*	*xk-at-wari*
3SG	*x-wari*	*n-wari*	*xti-wari*
1PL	*x-oq-wari*	*n-oq-wari*	*xq-oo-wari*
2PL	*x-ix-wari*	*n-ix-wari*	*xk-ix-wari*
3PL	*x-ee-wari*	*n-ee-wari*	*xk-ee-wari*

(Dayley 1985: 87–8)

(12.4) Kobon

	PRESENT	FUTURE	CONDITIONAL
1SG	*ar-ab-in*	*ar-nab-in*	*ar-bnep*
2SG	*ar-ab-ön*	*ar-nab-ön*	*ar-bnap*
3SG	*ar-ab*	*ar-nab*	*ar-böp*
1DU	*ar-ab-ul*	*ar-nab-ul*	*ar-blop*
2/3DU	*ar-ab-il*	*ar-nab-il*	*ar-blep*
1PL	*ar-ab-un*	*ar-nab-un*	*ar-bnop*
2PL	*ar-ab-im*	*ar-nab-im*	*ar-bep*
3PL	*ar-ab-öl*	*ar-nab-öl*	*ar-blap*

(Davies 1981: 166, 181)

Both these paradigms show zero expression in the third person singular. The Tzutujil paradigm shows that the non-indicative form (called 'potential') has a longer marker than the indicative forms, and the Kobon paradigm shows a longer marker for future tense than for present tense. The conditional mood in Kobon is marked by the two consonants *b* and *p*, so it is longer than the present indicative form, which has just a single consonant (this assumes that consonants are more important in counting length than vowels).

12.1.3 The correlation between frequency and differentiation

In three different senses, frequently used values tend to be more differentiated than rarely used values.[1] First, frequent values show *less syncretism* than rare values. Consider the partial paradigm of the Old English verb *bindan* 'bind' in (12.5).

(12.5)

	PRESENT IND	PRESENT SBJV	PAST IND	PAST SBJV
1 SG	*binde*	*binde*	*band*	*bunde*
2 SG	*bintst*	*binde*	*bunde*	*bunde*
3 SG	*bint*	*binde*	*band*	*bunde*
1–3 PL	*bindaþ*	*binden*	*bundon*	*bunden*

This paradigm shows that there is more syncretism in the plural than in the singular (in fact, all plural forms of all verbs are syncretized in Old English), more syncretism in the subjunctive than in the indicative, and more syncretism in the past indicative than in the present indicative. The same tendency is found in Khanty possessive suffixes:

[1] Generalizations about a correlation between frequency of use and linguistic structure are based on strong tendencies, but should not be treated as inviolable principles. For every generalization there are counterexamples.

(12.6)

	SINGULAR	PLURAL	DUAL
1ST	*-ēm*	*-ēw*	*-ēmən*
2ND	*-ēn*	*-lən*	*-lən*
3RD	*-l*	*-ēl*	*-lən*

(Nikolaeva 1999: 14)

Here, syncretism is found in the rarest of the three number values, the dual, and in one of the rarer person values, second person. (More syncretism in the dual can also be seen in Kobon (see (12.4) above)). More syncretism in the passive than in the active voice can be exemplified from Gothic (*niman* 'take').

(12.7)

	ACTIVE		PASSIVE	
	SINGULAR	PLURAL	SINGULAR	PLURAL
1ST	*nima*	*nimam*	*nimada*	*nimanda*
2ND	*nimis*	*nimiþ*	*nimaza*	*nimanda*
3RD	*nimiþ*	*nimand*	*nimada*	*nimanda*

The active has five different shapes, and the passive has only three.

The second sense in which frequent values are more differentiated is that inflection classes differ primarily with respect to the frequent values, less so with respect to rare values. In other words, the classes have *fewer shared* ***exponents*** in the frequently used values. This can be seen in Russian noun inflection. The endings of the four most important Russian inflection classes are shown in (12.8) (the inflection classes are labelled I–IV, as is traditional. See Exercise 5 of Chapter 8 for full word-forms belonging to each of these classes).

(12.8)

SINGULAR				
	IV	I	III	II
NOM	*-o*	Ø		*-a*
ACC				*-u*
GEN	*-a*		*-i*	
DAT	*-u*		*-i*	*-e*
LOC	*-e*			
INSTR	*-om*		*-ju*	*-oj*

PLURAL			
IV	I	III	II
-a	*-i*		
Ø	*-ov*	*-ej*	Ø
-am			
-ax			
-ami			

The contrast between singular and plural is clear: in the singular, there are at least twelve distinct endings, while in the plural there are at most eight. And, at least in the plural, the rarer cases (dative, locative, instrumental) show fewer allomorphs than the more frequent cases. Likewise in Standard

Arabic, transitive verbs belong to one of four inflection classes, characterized by different vowels before the final stem consonant. However, in the rarer passive voice the inflection is uniform and the difference between the inflection classes disappears (see (12.9)).

(12.9)

	ACTIVE		PASSIVE		
	PERFECT	IMPERFECT	PERFECT	IMPERFECT	
a-u:	*qatala*	*yaqtulu*	i-a: *qutila*	*yuqtalu*	'kill'
a-i:	*ḍaraba*	*yaḍribu*	i-a: *ḍuriba*	*yuḍrabu*	'hit'
i-a:	*ħafiẓa*	*yaħfaẓu*	i-a: *ħufiẓa*	*yuħfaẓu*	'protect'
a-a:	*jamaʕa*	*yajmaʕu*	i-a: *jumiʕa*	*yujmaʕu*	'gather'

The third sense in which frequently used values are more differentiated is that they tend to show *more cross-cutting values*. For example, as we saw in Section 5.1, the Latin future tense lacks a subjunctive mood (or one could alternatively say that the subjunctive mood lacks a future tense). In (12.10), we again see the third person singular of the verb *laudare* 'praise'.

(12.10)

	PRESENT TENSE	PAST TENSE	FUTURE TENSE
INDICATIVE	*laudat*	*laudabat*	*laudabit*
SUBJUNCTIVE	*laudet*	*laudaret*	—

Lack of cross-cutting values is similar, but not identical to syncretism. In Latin, the distinction between indicative and subjunctive is not neutralized in the future tense. The form *laudabit* ('she will praise') expresses only the indicative, and future tense cannot be expressed directly in the subjunctive.

12.1.4 Local frequency reversals

Table 12.1 shows the frequency asymmetries that hold in general in languages. However, in particular lexemes, the frequency relations may be reversed. For instance, while most nouns (such as 'table', 'head' or 'doctor') occur more often in the singular than in the plural, a small group of nouns tend to occur more often in the plural in many, if not all, languages. These are nouns referring to some paired or multiple body parts ('eyes', 'lips', 'hair(s)'), small animals ('ants', 'fish', 'mice'), small parts of plants ('beans', 'strawberries', 'leaves'), and some others ('sand grains', 'splinters').

In the case feature, nouns that denote a place occur in the locative case more often than in the nominative, in contrast to other nouns. And, while the greater relative frequency of the nominative case is clearly true of animate nouns that may occur as subjects of transitive clauses, it is not so clear that inanimate nouns, which are typically patients, are also used more frequently in the nominative than in the accusative case.

Local frequency reversals may also be found in particular cross-cutting

values. While in general the third person is more frequent than the second person, in the imperative mood this relation is reversed: commands are more often addressed to the person who is supposed to carry them out, and indirect imperatives (with the subject in the third person) are rare in all languages.

Structural effects of these frequency reversals can be observed in many languages. In Welsh, plurals are normally marked by suffixes as in other Indo-European languages (*cath/cathod* 'cat/cats', *draenog/draenogod* 'hedgehog/hedgehogs'; see (8.10) for more examples). However, in certain nouns that are used frequently in the plural, it is the singular that is marked by a special suffix:

(12.11)

dail	'leaves'	*deilen*	'leaf'
pysgod	'fish (PL)'	*pysgodyn*	'fish (SG)'
ffa	'beans'	*ffäen*	'bean'
cacwn	'wasps'	*cacynen*	'wasp'
mefus	'strawberries'	*mefusen*	'strawberry'
tywys	'corn'	*tywysen*	'ear of corn'

(King 1993: 67–9)

There are also languages in which case marking is found only in inanimate nouns (or non-personal pronouns). Godoberi is such a language, and here it is the (transitive) subject case that is overtly marked, whereas the direct-object case is zero (12.12).[2] Presumably, this is connected to the fact that inanimate nouns are more likely to be used as objects than as subjects.

(12.12)

(transitive) subject case	*den-Ø*	'I'	*hanqu-di*	'house'
direct-object case	*den-Ø*	'me'	*hanqu-Ø*	'house'

(Kibrik 1996: 119, 36)

And in the imperative, the second person form is often zero while the third person form is overtly marked (e.g. Latin second person imperative *lauda* 'praise!', third person imperative *laudato* 'let him/her praise!').

Local frequency reversals occur in derivational morphology as well. In many languages female person nouns are derived by a special affix from the corresponding male or general person noun – e.g. Dutch *handelaar* '(male) merchant', *handelaarster* 'female merchant', Hausa *àbookii* '(male) friend', *àbookìyaa* 'female friend'. From the point of view of the semantics, it would be equally possible to have a special affix that denotes male persons, but such affixes seem to be extremely rare. One reason for this asymmetry is probably that, in most societies, men tended to have more specialized

[2] Instead of the familiar terms 'nominative/accusative', the terms subject case/object case are used here, because overtly marked subject cases are usually called 'ergative' rather than 'nominative'.

roles, so that at least person nouns that denote professions and occupations are more frequently applied to men. Thus, the direction of derivation (from male/general to female) is related to frequency of use. However, the frequency relations tend to be reversed with nouns like *nurse* (because more women are nurses than men) and *widow* (probably not because husbands die more often than wives, but because marital status has traditionally been considered more relevant for women than for men). As a result of the unusual frequency relations, we get unusual male forms with overt marking (*widow-er, male nurse*).

12.1.5 Explaining the correlations

The correlation between frequency and shortness is clearly motivated by language users' preference for economical structures. Speakers can afford shorter expressions (or even zero expressions) when these are frequent, because frequent expressions are more predictable and are therefore those that are expected by default. The basic principle here is the same as in many other areas of human communication. For instance, in many countries local phone calls do not require an area code because phone calls to the local area are more common than phone calls to other areas.

In language, such economical structures may arise when a new distinction is introduced that is coded only in one of the two contrasting values. For instance, Spanish has a new subject/object distinction, which is marked by the preposition *a* with animate NPs (e.g. *Veo a mi hermano* [see.1SG to my brother] 'I see my brother'). This does not have morphological status yet, but if it becomes grammaticalized as an accusative case prefix, we will have a case system that conforms to the pattern in (12.12), in which the less frequently used case form gets the overt marking. The nominative was never marked overtly from the beginning of this change.

Another way in which an economical case-marking system may arise is by selectively preserving older markers. For example, in the Old High German *n*-declension, animate and inanimate nouns alike had a distinction between nominative and accusative (see (12.13)).

(12.13)

	Old High German		Modern German	
NOM.SG	*affo*	*knoto*	*Affe*	*Knoten*
ACC.SG	*affon*	*knoton*	*Affen*	*Knoten*
	'ape'	'knot'	'ape'	'knot'

Then the nominative/accusative distinction was lost in inanimate nouns, and in Modern German only animates preserve the zero marking in the nominative. Again, the resulting pattern conforms to (12.12), but it has come into existence via a different diachronic route.

The correlation between frequency and differentiation is due to the

greater **memory strength** of frequent values. When a value occurs rarely, it is more difficult to remember all the details of that value, so syncretism is more common in rare values, and various suppletive allomorphs are more easily kept apart in the frequent values.

12.2 The direction of analogical levelling

Analogical levelling is a common type of morphological change. Levelling eliminates morphophonological alternations by extending one stem alternant to other word-forms in the paradigm. For instance, many speakers of English have eliminated the alternation in *house/houses*, which in the traditional pronunciation has a voiced final stem consonant in the plural: [haʊs]/[haʊzəz]. Now crucially, it is the form of the singular stem that is extended by the innovating speakers ([haʊs]/[haʊsəz]), not the plural stem. There are no English speakers that pronounce the singular noun *house* as [haʊz].

This change is typical of analogical levelling in general: the form of the stem that is extended within the paradigm is usually the value with higher frequency. That frequency is the crucial factor is particularly clear from cases of local frequency reversals. A particularly striking case of this comes from West Frisian, where in the traditional language many nouns show a vowel alternation in singular–plural pairs. In innovative varieties of the language, this alternation is eliminated and the singular and plural stems are identical again, (see (12.14)).

(12.14)

	conservative	innovative	
a.	*hoer/hworren*	*hoer/hoeren*	‘whore(s)’
	koal/kwallen	*koal/koalen*	‘coal’
	miel/mjillen	*miel/mielen*	‘meal, milking’
	poel/pwollen	*poel/poelen*	‘pool(s)’
b.	*earm/jermen*	*jerm/jermen*	‘arm(s)’
	kies/kjizzen	*kjizze/kjizzen*	‘tooth/teeth’
	hoarn/hwarnen	*hwarne/hwarnen*	‘horn(s)’
	trien/trjinnen	*trjin/trjinnen*	‘tear(s)’

(Tiersma 1982: 834)

In (12.14a), the singular stem form is extended in analogical levelling, but, in (12.14b), the plural stem form is extended. The choice of the form that is extended is by no means arbitrary: when the noun denotes a thing that tends to occur in groups and hence is more frequent in the plural, the plural stem wins out.

An example from case inflection is Latin *oleum* ‘olive tree’, which goes back to an earlier form *oleivum* (cf. *oleiva,* later *olīva,* ‘fruit of the olive tree,

olive'). Then three sound changes occurred: (i) the diphthong *ei* turned into *ē* and later into *ī*, (ii) the semivowel *v* [w] was dropped before *u* and (iii) long vowels were shortened before another vowel. As a result, the nominative/accusative form *oleivum* successively became *olēvum*, *olēum* and *oleum*, whereas the genitive and dative forms *oleivī/oleivō* became *olīvī/olīvō*. Then, analogical levelling extended the nominative/accusative stem to the other case forms (*oleiva* became *olīva* and retained the stem *olīv-*, because the *v* never dropped from its paradigm):

(12.15)

	oldest form	later form	Classical Latin
NOM/ACC.SG	*oleivum*	*oleum*	*oleum*
GEN.SG	*oleivī*	*olīvī*	*oleī*
DAT.SG	*oleivō*	*olīvō*	*oleō*

The greater stability of frequent stem forms can be explained again by memory strength and speed of lexical access. The genitive singular *olīvī* is replaced by *oleī* because the stem *ole-* has higher memory strength and may thus be used when a speaker (temporarily) forgets the old form *olīv-*, or because *ole-* can be retrieved more quickly from the lexicon and combined with the suffix *-ī* than the form *olīvī*, with its rarer stem form *olīv-*.

12.3 Frequency and irregularity

In language after language, if there are irregularities in inflection, these primarily affect the most frequent lexemes. Our first example comes from Koromfe, which has scores of regular verbs like those in (12.16a), and a few irregular verbs like those in (12.16b).

(12.16)

a.

HABITUAL	PAST	
kam	*kamɛ*	'squeeze'
tarı	*tarɛ*	'plaster'
leli	*lele*	'sing'

b.

HABITUAL	PAST	
bɛ	*bɛn-ɛ*	'come'
bo	*bol-e*	'say'
tɛ	*tɛr-ɛ*	'arrive'

(Rennison 1997: 271–5)

In Welsh, there are four irregular verbs whose past tense is totally unlike the past tense of a regular verb such as *gwel-* in (12.17a). Three of them are shown in (12.17b).

(12.17)

	a. *gwel-d* 'see'	b. *myn-d* 'go'	*gwneu-d* 'do'	*do-d* 'come'
1SG	*gwel-es i*	*es i*	*nes i*	*des i*
2SG	*gwel-est ti*	*est ti*	*nest ti*	*dest ti*
3SG	*gwel-odd e*	*aeth e*	*naeth e*	*daeth e*

(King 1993: 183)

In Old English, grammars list just four verbs that are totally irregular and cannot be fitted into any of the inflectional classes. These are shown in (12.18b), and a regular verb is shown in (12.18a).

(12.18)

	a. 'bind'	b. 'be'	'do'	'go'	'want'
1SG.PRS	*binde*	*eom*	*dō*	*gā*	*wille*
2SG.PRS	*bintst*	*eart*	*dēst*	*gǣst*	*wilt*
3SG.PRS	*bint*	*is*	*dēþ*	*gǣþ*	*wille*
1–3PL.PRS	*bindaþ*	*sint*	*dōþ*	*gāþ*	*willaþ*
1SG.PST	*band*	*wæs*	*dyde*	*ēode*	*wolde*
PARTICIPLE	*gebunden*	—	*gedōn*	*gegān*	—

Thus, the verbs that tend to show irregularities are those that mean 'be', 'do', 'go', 'come', 'say', and so on – i.e. precisely those verbs that are used the most frequently across languages.

In nouns, the situation is more or less the same. For example, in Lango regular plural suffixes are *-ê, -nì* and *-í.* Some regular and most of the irregular nouns are listed in (12.19).

(12.19)

a.				b.			
	réc	*réc-ê*	'fish(es)'		*dákô*	*món*	'woman/women'
	pʊ́nʊ̂	*pùn-nì*	'pig(s)'		*ɲákô*	*àɲìrà*	'girl(s)'
	lɛ̀	*ley-í*	'axe(s)'		*ìcɔ̀*	*cɔ̀*	'man/men'
					dánô	*jɔ̀*	'person/people'
					dyàŋ	*dòk*	'cattle'
					gìn	*gìgù*	'thing(s)'

(Noonan 1992: 83–5)

Irregular noun plurals in Bulgarian include *oko/oči* 'eye(s)', *uxo/uši* 'ear(s)', *dete/deca* 'child(ren)', and Italian has the three irregular nouns *uomo/uomini* 'man/men', *dio/dei* 'god(s)', *bue/buoi* 'ox(en)'. The appearance of words for 'cattle' and 'ox' on several of these lists may at first seem surprising – these are certainly not among the most frequent nouns in modern Italian and modern English. But in modern Lango they may well be (cattle herding is one of the main economic activities of Lango speakers), and in older Italian and older English the situation may have been similar.

There are two rather different ways in which frequency may cause irregularity in morphology. On the one hand, frequency leads to phonological reduction, because frequent expressions are relatively predictable, so that speakers can afford to articulate less clearly. This factor must be invoked to explain the irregularities in Koromfe verbs in (12.16). Examples from English are the verbs *have*, *say* and *make*, which were completely regular in earlier English (*haved, sayed, maked*), but became irregular because they were subjected to greater phonological reduction than comparable rarer verbs (e.g. *said* versus *played*, *had* versus *behaved*, *made* versus *faked*).

On the other hand, frequency leads to memory strength and fast lexical access, so that frequent items are less susceptible to analogical levelling and other regularizations. So, while frequency causes faster phonological change, with respect to morphology it has a conserving, decelerating function. For example, the irregular Italian noun *uomo/uomini* 'man/men' preserves an old declension type inherited from Latin (*homo/homines*) that was otherwise eliminated by regularizing changes (cf. Latin *virgo/virgines* 'virgin(s)', Italian *vergine/vergini*). This conserving effect of frequency is also the cause of the Bulgarian irregular plurals *oči* 'eyes' and *uši* 'ears'. These were originally dual forms, and, because eyes and ears typically occur in pairs, these word-forms were probably the most frequent forms in the paradigm. Since eyes and ears are among the most frequently used paired body parts, it is not surprising that these forms survive.

From a diachronic point of view, the least well-understood type of irregularity is stem suppletion, as seen in Welsh *myn-/es-/aeth*, Old English *is/wæs*, *gǣþ/ēode*, and Lango *dákô/món*. It is difficult to understand why speakers would begin to associate roots that originally came from two different lexemes and integrate them as word-forms of the same lexeme. But, granted that speakers sometimes do that, the conserving effect of frequency will maintain the suppletion in the most frequent lexemes. It is also worth pointing out that inflection class differentiation (which we discussed in Section 12.1.3) works in exactly the same way: different markers for the same meaning/inflectional values can be maintained if the items affected are sufficiently frequent, whether owing to the frequency of the inflectional value or to lexeme frequency.

Summary of Chapter 12

Token frequency is relevant to morphology because frequent words occur more predictably in context, are more easily remembered and are retrieved faster than rare words. Because speakers favour economical structures, the greater predictability of frequent values typically results in zero expression (or otherwise short expression). Frequent values are also more differentiated – they show less syncretism, fewer shared exponents and more cross-cutting values. Because frequent words and values are more easily remembered, they are less subject to analogical levelling, and this is also one of the reasons why irregularities exist mostly in frequent words. Another reason is that frequent words are subject to greater phonological reduction, again because of predictability. Over time, frequency effects thus shape (inflectional) morphological structure in a number of ways.

Further reading

Frequency differences between inflectional values of the same feature are discussed (under the name of 'markedness') by Greenberg (1966) and Croft (1990: ch. 4). Haspelmath (2006) argues that an abstract notion of markedness is superfluous once the role of frequency is appreciated.

The insight that frequency is the explanation for shortness was already emphasized by Zipf (1935). For local frequency reversals, see Tiersma (1982). For the relation between frequency and irregularity, see Mańczak (1980a, b), Werner (1989), Bybee (1995), Nübling (2001), Corbett *et al*. (2001) and Brown *et al.* (2007). For the view that grammatical structure (including morphology) cannot be adequately understood without considering frequency effects, see Bybee (2006) and the papers in Bybee and Hopper (2001).

Comprehension exercises

1. The general correlation between frequency and shortness leads to certain expectations about inflectional paradigms. Consider the following (partial) paradigms and determine where these expectations are fulfilled, and where we should be surprised.

a. Udmurt conjugation: past tense of *uć-* 'look'

1SG	*ućki*	1PL	*ućkimy*
2SG	*ućkid*	2PL	*ućkidy*
3SG	*ućkiz*	3PL	*ućkizy*

(Perevoščikov 1962: 203)

b. Even declension: *ǰuu* 'house'

	SG	PL
NOM	*ǰuu*	*ǰuul*
ACC	*ǰuuw*	*ǰuulbu*
DAT	*ǰuudu*	*ǰuuldu*
COM	*ǰuuñun*	*ǰuulñun*
ABL	*ǰuuduk*	*ǰuulduk*

(Malchukov 1995: 9)

c. Pipil possessive inflection: *nu-chi:l* 'my chilli pepper', etc.

1SG	*nu-chi:l*	1PL	*tu-chi:l*
2SG	*mu-chi:l*	2PL	*amu-chi:l*
3SG	*i-chi:l*	3PL	*in-chi:l*

(Campbell 1985: 43)

d. Tauya possessive inflection: *ya-potiyafo* 'my hand', etc.

1SG	*ya-potiyafo*	1PL	*sono-potiyafo*
2SG	*na-potiyafo*	2PL	*tono-potiyafo*
3SG	*potiyafo*	3PL	*nono-potiyafo*

(MacDonald 1990: 129–30)

2. The Modern French verb *trouver* 'find' used to have two different forms of the stem in older French, *trouv-* and *treuv-*. The former occurred in word-forms that were stressed on the suffix, and the latter occurred in word-forms that were stressed on the stem. (A dot below the syllable indicates the position of the stress.) This stem alternation no longer exists in modern French: all forms of the verb *trouver* have the same stem vowel. Why is this change surprising after what we learned in this chapter?

	older French	modern French
'I find'	*je trẹuve*	*je trọuve*
'you find'	*tu trẹuves*	*tu trọuves*
'he finds'	*il trẹuve*	*il trọuve*
'we find'	*nous trouvọns*	*nous trouvọns*
'you(PL) find'	*vous trouvẹz*	*vous trouvẹz*
'they find'	*ils trẹuvent*	*ils trọuvent*

3. Go back to Chapter 10, where morphophonological alternations were discussed. Where did we make reference to frequency in that chapter? How did what we said there fit with the claims of this chapter?

Exploratory exercise

In this chapter we saw that frequency asymmetries and structural asymmetries are correlated. For instance, a noun lexeme's singular forms tend to be more frequently used than its plural forms, and correspondingly, case forms tend to be more differentiated in the singular and singular exponents tend to be shorter than their plural counterparts. This pattern can be seen even at the level of individual lexemes. Where the plural is more frequently used (a frequency reversal), that lexeme is likely to exhibit more differentiation in the plural, and shorter plural forms.

We also claimed that frequent lexemes tend to be irregular. The reader may have noticed, however, that this discussion was based on a different type of frequency comparison. Rather than looking at the relative frequency of different paradigm cells, we compared the frequency of use of different lexemes. We did not consider whether irregularity correlates with frequency asymmetries within a lexeme. But given the importance of frequency at this level, we might ask whether such a correlation exists. In other words,

are more frequently used cells in a paradigm more (or less) likely to be irregular? You will develop an answer to this question.

This exercise is based on (but simplifies) Corbett *et al.*'s (2001) study of Russian, and we use Russian for demonstration purposes below. You need not choose this language to investigate, but a good frequency dictionary or frequency list, with information for individual word-forms, must be available. (Hint: to make the process of finding word-form frequencies more efficient, it is helpful if the frequency dictionary/list is available for electronic searching.)

Instructions

Step 1: Choose a morphological pattern that exhibits (strong) stem suppletion in at least a handful of lexemes. A few examples of Russian singular–plural noun pairs are given below. As can be seen, the examples in (12.20a) have the same stem for the singular and plural, but the ones in (12.20b) exhibit suppletion.

(12.20)

	SINGULAR	PLURAL	GLOSS
a.	*zavod*	*zavody*	'factory'
	student	*studenty*	'student'
b.	*syn*	*synov'ja*	'son'
	rebënok	*deti*	'child'
	čelovek	*ljudi*	'person'

Since suppletion in the Russian examples is according to number, we want to ask how frequently the plural is used, compared to the singular, and whether this differs depending on whether the noun is regular or suppletive. An appropriate measure is thus the ratio of plural frequency to singular frequency, e.g. token frequency of *ljudi* divided by token frequency of *čelovek*. (For demonstration purposes we are ignoring the fact that Russian has nominal cases.)

Step 2: Develop a hypothesis. Based on what you have read in this chapter and elsewhere in the book (e.g. discussion of frequency in Section 4.3), make a guess about what the answer to the research question will be. For instance, would you expect Russian lexemes with suppletive stems in the plural to be more/less frequently used in the plural than in the singular? What would you expect for lexemes with the same stem throughout? Explain your reasoning.

Step 3: Build a list of lexemes with stem suppletion. Identify as many relevant words as possible.

Step 4: Using a frequency dictionary or frequency list, gather token frequency counts. For each of the suppletive lexemes from Step 3, find the token frequency of each word-form in its paradigm. Then, do the same for at least 10 lexemes that have the same stem throughout the paradigm (the

(12.20a) type), and which are of similar overall frequency to the suppletive lexemes (or as close as possible).

Step 5: For each lexeme, calculate the relevant frequency ratio. For instance, according to one count, the lexeme ČELOVEK occurs in the singular 1678 times per million words of text, and in the plural 1267 times per million words of text (Sharoff 2002). Its ratio is thus 0.755 (= 1267/1678).

Step 6: Evaluate the data and draw conclusions. Compare the relative frequency ratios for the two groups. Are there any notable differences in the frequency ratios? Do the results match your predictions? Consider the implications for understanding the relationship (or lack thereof) between the frequency of paradigm cells and irregularity.

Alternative: Irregularity can be treated as a scale ranging from strong suppletion to full regularity (see Chapter 2). Rank different kinds of stem irregularity, and look for a correlation between *degree* of irregularity and frequency. (This is much harder!)