13.1 Introduction

When two languages are found to share a grammatical property P, there are four possible reasons: (a) it is accidental, (b) they have a common ancestor language, (c) they have been in contact and one language has taken over P from the other language, (d) P is universal, i.e. all languages have P, not because they all derive from a common source but by necessity, because of ‘the nature of language’, or in other words, P is determined by universal grammar. The notion universal grammar was introduced into modern linguistics by Noam Chomsky in the early sixties, as part of the generative research programme (Chomsky 1965:5-8). The idea that all human languages are special instances of a universal format of language has a long history, though, including Roger Bacon and the modistae (or speculative grammarians) of the thirteenth and fourteenth centuries, as well as the Port Royal grammarians and the philosophical grammar movement in the seventeenth century, which were particularly important for Chomsky’s own project (see Chomsky 1968:12-19).¹

¹ Wikipedia’s entry on universal grammar mentions that the article on ‘Grammar’ in the first edition of the Encyclopædia Britannica (1771) contains an extensive section entitled ‘Of Universal Grammar’.
The way the idea developed in works by Chomsky and other generative linguists at the time, and the way it came to be understood, is as an innate, cognitive faculty which is crucial for the acquisition and use of human language (see chapter 18, this volume). Behind all the variation exhibited by the languages of the world, there would be formal properties that are common to all languages because they are determined by genetically encoded properties of our brain, a product of the biological evolution of our cognitive system.

The notion that there are universal grammatical properties has been advanced on two fronts, though, ever since the early sixties: within generative linguistics and within linguistic typology, emanating from the work of Joseph Greenberg. In linguistic typology there is no commitment to the idea that the universals are ultimately genetically encoded (for discussion, see chapter 35, this volume). They are generalizations; their explanation may vary from case to case.

In this chapter I will keep distinct two senses of universal grammar. One is universal grammar in the broad sense, which is all the grammatical properties that are characteristic of human languages, regardless of why they are there. They may be the product of cultural or biological evolution (see chapter 32, this volume). The other is universal grammar in the strict ‘Chomskyan’ sense, which is the grammatical properties found in all human languages that are ultimately a product of biological evolution. Following tradition, I will refer to the latter sense of universal grammar as UG. I will focus mainly on UG, and the development that this theoretical construct has undergone ever since its inception in Chomsky’s work in the sixties. As will be shown, it has undergone some considerable changes in the course of its lifetime. As will also be shown, the contribution of Noam Chomsky has been crucial at every turn
in this development. I will begin, though, by a short review of the research on
universals within linguistic typology.

13.2 Linguistic typological universals

Much of the evidence for universal grammar in the broad sense comes from cross-
linguistic, comparative research. If similarities among languages are discovered
which are unlikely to be accidental, and cannot be explained by common origin (cf.
chapter 9, this volume) or language contact (cf. chapter 8, this volume), then, by
definition, they belong to universal grammar in the broad sense. A very large of
number of such similarities have indeed been discovered.\(^2\) One line of comparative
research with the stated aim of discovering linguistic universals is the typological one
(for full discussion, see also chapter 35, this volume), starting with the work of
Joseph Greenberg in the early sixties. On the basis of a sample of 30 languages from
a variety of families and locations, Greenberg (1963) proposed a set of 45 cross-
linguistic generalizations which he termed universals of language. They are

\(^2\) The Konstanz Universal Archive lists about 1400 linguistic universals. This list
includes all universals that have been proposed in the literature, regardless of how
hypothetical the basis for them, and regardless of whether they have since been
shown to be spurious. It includes not just absolute universals but also ‘statistical
universals’, that is more or less strong tendencies, following in Greenberg’s
footsteps. It does not include many universals proposed within the generative
tradition, perhaps because they are not typically explicitly referred to as ‘universals’
in the written sources, even though they are typically assumed, by default, to be
universal.
sometimes termed typological universals, to distinguish them from the linguistic universals assumed in the generative tradition. As mentioned, in the case of Greenberg there is no commitment to the idea that the universals are ultimately genetically encoded. The epithet ‘universal’ has no other significance than that a given generalization, as far as is known, holds true of all languages or ‘the overwhelming majority of languages’; in either case there is something to explain. The favoured method within this research programme is comparison of a large set of languages, preferably hundreds, with respect to some grammatical property, where the languages are sampled so as to represent all the major language families and all different corners of the world, in order to avoid genetic bias or areal bias. In practice this means that the data are collected mainly from existing grammatical descriptions of languages.

Within the Greenbergian typological research programme as developed by Lehmann (1973; 1978), Venne\neman (1974; 1984), Hawkins (1979; 1983), Greenberg, Ferguson and Moravcisk (1978), and later by the linguists associated with the World Atlas of Language Structures (WALS) programme (Bernard Comrie, Martin Haspelmath, Matthew Dryer, David Gil, among many others), there has been a strong tendency to favour functional over formal explanations of typological universals, where a functional explanation of a grammatical property P is typically that P makes communication by language more efficient, and therefore all languages have evolved, by cultural evolution, to encompass P; see Newmeyer (1998), and chapter 36, this volume. A different type of explanation, still functional, would be

---

3 Bill Croft could also be mentioned as an influential theorist in the typological camp (cf. Croft 1995; 1999).
that P makes language acquisition more efficient, and therefore languages have evolved to encompass P; see Kirby and Hurford (2002), Kirby and Christiansen (2003) for a particular version of this idea. In Holmberg (to appear) I suggest that there is no logically necessary connection between linguistic typology and functionally-based explanation, but that there is a practical connection to do with the methodology. I quote:

The favoured method in linguistic typological research is comparative surveys on a large scale in order to cover as much as possible of the existing variation and in order to establish, as far as possible, valid global generalizations. This has meant that the grammatical properties that are investigated/compared, are by necessity all easily observable ‘surfacy’ properties of the kind which are recorded even in sketchy descriptive grammars. One result of this is that the generalizations discovered have been probabilistic, riddled with exceptions, rather than absolute, because surfacy properties are subject to unpredictable variation to a greater extent than more abstract properties […] This disfavours explanations in terms of universal, genetically determined properties of the language faculty, and favours explanations in terms of ‘functional pressure’, which are expected to allow for exceptions.

Holmberg (to appear).

Among the universals in Greenberg (1963) which have generated most discussion and most new research are the so called word order universals (universals 1 to 25). A subset of them are given here, retaining the numbering in Greenberg (1963):
3. Languages with dominant VSO order are always prepositional.

4. With overwhelmingly greater than chance frequency, languages with SOV order are postpositional.

5. If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun.

9. With well more than chance frequency, when question particles or affixes are specified in position by reference to the sentence as a whole, if initial, such elements are found in prepositional languages, and, if final, in postpositional.

13. If the nominal object always precedes the verb, then verb forms subordinate to the main verb also precede it.

16. In languages with dominant order VSO, an inflected auxiliary always precedes the main verb. In languages with dominant order SOV, an inflected auxiliary always follows the main verb.

These cross-linguistic generalizations have the fascinating property of indicating an underlying pattern, as was observed by Greenberg and formally described in terms of the notions ‘harmony’ and ‘dominance’. Subsequent research led to formulation of the pattern in terms of head-complement order: There is a tendency for phrases to be either head-initial or head-final, across categories; see Dryer (1992). There is a statistical correlation between (S)OV, [NP P], [NP N], [VP Aux], [IP C] on the one hand, and VO (VSO or SVO), [P NP], [N NP], [Aux VP], and [C IP] on the other.

Dryer (1992) opted for formulation of the generalization in terms of branching direction: languages tend to be either left-branching or right-branching across categories.
hand, even though many, (even most), languages exhibit a mixture of head-initial and head-final phrases.⁵

Greenberg’s word order correlations were subsequently tested and, for the most part, confirmed, against a much bigger and more carefully sampled set of 625 languages in Dryer (1992). Dryer’s research also demonstrated that some of Greenberg’s universals were spurious, being effects of language contact, including Universal 5; see Dryer (1988). The WALS project (Haspelmath, Dryer, Gil and Comrie 2014) has also shown some of Greenberg’s universals to be spurious. In spite of these minor objections, Greenberg’s research and subsequent typological research has shown persuasively that there is unity, simplicity and some sort of rationality underlying the seemingly unconstrained surface variation, which obviously calls for explanation.

I will come back to the typological universals, after a discussion of the development of universal grammar in the strict sense (UG) in generative linguistic theory.

13.3 Arguments for UG (strict sense)

A particularly compelling argument for UG in the strict sense, first expounded in Chomsky (1965) is the argument from ‘the poverty of the stimulus’. Insofar as it can be shown that speakers of a language have intuitions about interpretation and grammaticality of constructions that are so complex and/or so rare that the intuitions

⁵See Baker (2008) for a demonstration that the correlation between head-complement order across categories holds as a universal tendency, in spite of the prevalence of mixed systems.
cannot reasonably be the result of learning purely from experience, i.e. just from exposure to relevant data in conversation, then the intuitions must be based on properties of our cognitive system that are in some sense innate.

The following is an example of this form of argument, from Chomsky (1986). Consider the sentences (1)-(4).

1 John ate an apple.

2 John ate.

3 John is too stubborn to talk to Bill.

4 John is too stubborn to talk to.

I quote from Chomsky (1986); the example numbers are changed.

[...] sentence (1) means that John ate something or other, a fact that one might explain on the basis of a simple inductive procedure: *ate* takes an object, as in (2), and if the object is missing, it is understood as arbitrary. Applying the same inductive procedure to (3) and (4), it should be that (4) means that John is so stubborn that he (John) will not talk to some arbitrary person, on the analogy of (2). But the meaning is, in fact, quite different: namely, that John is so stubborn that some arbitrary person won't talk to him (John). Again, this is known without training or relevant evidence. (Chomsky 1986:8).
The conclusion is that the interpretation is based on intuitive knowledge of abstract principles of syntax that are inherent to the language faculty, not learnt but only triggered by experience.

One may or may not agree with the premises or the conclusion in the case of this particular example; here, it serves as an illustration of this form of argument. See Berwick, Pietroski, Yankama and Chomsky (2011) for some recent discussion with many more examples.

The argument from the poverty of the stimulus is particularly persuasive when it can be demonstrated that children at a very early age have the relevant intuitions (see also discussion in chapter 18, this volume). The following example is from de Villiers, Vainikka and Roeper (1990).

5  When did the boy tell his father that he hurt himself?
6  When did the boy tell his father how he hurt himself?

Sentence (5) is ambiguous: *when* can be about the time of telling (the short-distance reading) or the time of hurting himself (the long-distance reading). (6) is unambiguous: the long-distance reading is unavailable. The reason for this is the intervening question word *how*: the long-distance reading requires movement of *when* out of a clause headed by a *wh*-phrase, which violates the *wh*-island constraint, a universal constraint on movement (Ross 1967; Chomsky 1977) which will be discussed in a little more detail below. Between the ages of three and six, children give long distance interpretations to questions such as (5) roughly half of the time, but only around 6% for questions such as (6) (de Villiers, Roeper and Vainikka 1990; Pérez-Leroux 1993; Roeper and de Villiers 1994).
It is very unlikely that the children could have acquired these intuitions from experience, i.e. by exposure to relevant data. First, long distance questions are uncommon in ordinary discourse. Second, learning on the basis of exposure to data that (2) does not have a long-distance reading would require exposure to relevant negative data, that is data which indicate that the relevant reading is unavailable. But ordinary discourse does not include such data.\(^6\) If it is not learnt by experience, the relevant intuitions must be based on constraints or principles that are in some sense innate. As such they may be specific to the language faculty or they may derive from more general constraints on cognition. The methodological principle which has been assumed, quite rightly, within generative grammar is that insofar as we are unable to show that the properties have any application outside language, we are entitled to assume that they are specific to the language faculty; see Berwick, Pietroski, Yankama and Chomsky (2011).

13.4 The content of UG

Chomsky (1965), in the first detailed discussion of UG, proposed making a distinction between substantive and formal universals. Among the substantive universals would be categories characteristic of language: in phonology, vowels and

\(^6\) Negative evidence does occur: If a form can be expected to occur with some frequency in a language L, given general principles and given what is already known by a learner about L, but systematically fails to occur, this may well serve as indirect negative evidence. However, long distance wh-movement is surely not frequent enough to provide the required indirect negative evidence needed to learn the contrast between (1)-(2) by experience.
consonants and features such as front vs back, roundedness, nasality, etc.; in syntax, nouns, verbs, adjectives, adpositions, pronouns, negation, tense, aspect, number, etc. Some of these categories are universal in the sense that all languages have them. There is every reason to think that all languages have negation, for example. One of Greenberg’s (1963) universals states that all languages have a category of pronouns distinguishing at least three persons and two numbers (see Daniel 2011). But for many of the categories it is controversial whether they occur in all languages. For example, many languages appear not to have adjectives as a category distinct from nouns or verbs (but see Baker 2003, Dixon 2004 for arguments that adjectives are universal), many languages do not have any overt marking for tense (Dahl and Velupillai 2011; Ritter and Wiltschko 2011), many languages do not have any marking for nominal plurality outside the pronominal system (Daniel 2011), and so on. It is not obvious that all languages make even the traditional distinction between noun and verb; see Baker (2003) for discussion.

A widely entertained hypothesis is that UG makes available a pool of categories, from which each language makes a selection. This is the line taken by Chomsky (1995:ch. 4; 2001), for example. Another hypothesis is that all the categories and features are universal, present in all languages, and that the variation is a matter of which of them have morphological realization; see Cinque (1999), Sigurðsson (2004; 2011a,b).

7 In a sample of 261 languages, Daniel (2011) found 9 that do not distinguish number in their independent pronoun system, so that part of Greenberg’s generalisation is too strong.
Formal universals are, for example, the fact that syntactic structure is hierarchical, that syntactic rules are recursive, that there is movement, that there is labelling (phrases being labelled by their head), and so on. Though some of them are found in all languages, being constitutive of the grammar of human language, there are other formal properties that are not instantiated in every language. For example, functional heads may be morphologically bound as affixes or clitics, but not all languages make use of this option. There is even some debate whether all languages make use of recursion in the sense of self-embedding (Everett 2005; 2007; Nevins, Pesetsky and Rodrigues 2009a,b; Rodrigues and Sândalo to appear).

Another property of UG proposed by Chomsky (1965) but not much elaborated in that work, is what he called an evaluation metric. This would be a set of principles, activated under language acquisition, which would select the most highly valued grammar compatible with the primary data. This is based on the idea that acquisition of a language L, particularly as a first language, is a matter of constructing the grammar of L on the basis of (a) UG and (b) primary data, by a process in which hypotheses are formed by the learner about the mapping between sound and meaning in L, which are then tested and refined in the light of more data, until the grammar of L, which is essentially a theory of L, is complete and able to generate the same range of sentences as the grammars of other people in the linguistic community. However, the formal devices made available by UG are powerful enough to allow for more than one theory which is compatible with the primary data. Movement is a particularly powerful device, in this regard. Consider the following, somewhat trivial example: The embedded question in (7) can be derived by movement of where from the PP to initial position, as shown in (8). It can also be derived by movement of the PP,
deriving (9a) as an intermediate structure, followed by movement of the preposition back into the VP, deriving (9b), which would be spelled out as (7).

7 (I wonder) where it came from.

8 where [s it [vp came [pp from __]]]

9 a [pp from where] [s it [vp came __]]
   b [pp __ where] [s it [vp came from]]

The evaluation metric would prefer the grammar which derives (7) by one movement instead of two: other things being equal, a grammar with shorter derivations is preferred. The evaluation metric may also disprefer the grammar deriving (7) by (9a,b) because it employs downwards movement. Other things being equal, a grammar with fewer formal devices, for instance only upwards movement, is preferred. The evaluation metric of Chomsky (1965) later reappeared in minimalist theory as economy conditions, as will be discussed below in section 8 (see also discussions in §7.3.1, §24.4). It also reappeared much sooner, beginning already in the late sixties with Ross (1967) as a forerunner, as conditions on movement, to be discussed in the next section.

13.5 From conditions on transformations to principles and parameters theory

The Chomskyan theory of UG went through an interesting development in the seventies, culminating in Chomsky’s Lectures of Government and Binding, published in 1981. The first step was a reorientation from articulating the set of rules deriving
all and only the grammatical sentences of a given language, as in the models of Chomsky 1957, 1965, to articulating the formal properties of the rules and the representations derived by them as universal principles, so that the rules themselves could be correspondingly generalized and simplified.

This approach was pioneered in Ross (1967), where a set of constraints on movement were formulated, prohibiting movement out of certain types of constituents, called islands. Chomsky (1973) proposed some additional constraints on movement, in particular on what would later be identified as A-movement, and on pronominal and anaphoric binding, including (10), which explains, among other things, the contrast between (2a,b).

10  **The Tensed-S condition**: No rule can involve X,Y in a structure [...X... [\_
...Y...]], where \(\alpha\) is a tensed sentence.

11  a  *John seems \(\alpha\) that __ is tired

    b  John seems \(\alpha\) __ to be tired

Given (10) and certain other conditions like it, the movement rule itself needs only a very general formulation, at the limit simply ‘Move NP’.

The Tensed-S condition would be a property of UG. Given that the validity of this condition can be observed only on the basis of negative evidence (such as (11a)), and given that the required negative evidence does not occur in normal discourse, it cannot be learnt on the basis of linguistic experience, so the condition has to come from UG.
Chomsky (1977) was another important contribution in the evolution of the theory in the seventies. In this paper Chomsky showed that a wide variety of constructions, which up until then had been held to be quite distinct, each derived by a different set of rules, could be analysed as derived by one and the same rule, namely *wh*-movement, the application of which was subject to a locality constraint called Subjacency.

12 Subjacency: No rule can involve $X,Y$ in a structure $[...X[α [β ...Y ...]]$ where $α$ and $β$ are cyclic nodes.

   Cyclic nodes: S and NP (corresponding to IP and DP in more recent theory).

   ‘Cyclic node’ was later rephrased as ‘bounding node’. Subjacency is a generalization over a range of island constraints: The *wh*-island constraint, the complex NP constraint, the sentential subject constraint, and the adjunct island constraint, which were now seen to be instances of the same condition on movement.

   An important complement to the theory of *wh*-movement was Rizzi (1978) discussing the observation that *wh*-movement appears to allow violation of Subjacency in certain cases in Italian, as shown in (13).

13 a *Your brother, who I wonder which stories they have told to, is very worried.

   b
   Tuo fratello, a cui mi domando che storie abbiano raccontato, era
Rizzi proposed to account for this by assuming a partly different formulation for Subjacency in Italian: The bounding nodes would be CP and DP, not IP and DP, as in English. In the relative clause in (13b) the wh-phrase a cui has moved across two IP-boundaries, but only one CP boundary.

This pointed to a solution of a major problem for the reorientation from language specific rules to universal principles, namely how to account for cross-linguistic variation: a universal principle (Subjacency) allowed for variation on a certain point, namely, the identity of bounding nodes.\(^8\)

In the generative grammars of Chomsky (1957; 1965) and subsequent work grammatical variation was accounted for by the system of rules, which was different from language to language. Even though the rules were based on categories drawn from UG and had formal properties dictated by UG, the set of rules could vary indefinitely: a rule could be present or absent in a language, and the formulation of a rule could vary in many ways from language to language (as well as from one period

---

\(^8\) English and Italian may not, however, be as different as they are made out to be in Rizzi (1978); see Grimshaw (1986).
to another) as long as it observed the formal conditions set by UG. Variation was expected. At the same time, however, the system allowed too much variation, failing to capture some of the underlying unity of language, and thereby failing to address ‘the logical problem of language acquisition’ (also called ‘Plato’s problem’ by Chomsky 1986), namely, how it is possible to acquire such a complex system on the basis of so little evidence.

Once the properties of the constructions/rules were generalized as principles of UG, with the rules correspondingly simplified, the opposite problem arises, which is how to account for the abundant (synchronic and diachronic) cross-linguistic variation. As UG gets more specified, there is correspondingly less room for variation. However, if some principles of UG allow for variation on certain points, which came to be known as parameters, and if the principles are involved in a wide range of constructions, then even very limited variation in the grammar can have far-ranging effects on the language.

A full-blown version of this theory, which came to be known as principles-and-parameters theory (see chapters 7 and 31, this volume), was articulated in Chomsky (1981). The theory at that point encompassed an array of modules/subtheories, each defined by a set of principles, some of which allowed for parametric variation: theta-theory, case theory, binding theory, control theory, and bounding theory.

Another influential idea was that grammatical variation would all be lexical in the sense of being a matter of variation with respect to features of functional heads: Does I(NFL), the head of IP contain AGR, a set of nominal agreement features, or not? Does C have a feature attracting a verb or not, does definite D attract N or not, and so on. Borer (1984) is usually credited with articulation of this idea (which, incidentally, is not compatible with Rizzi’s (1978) subjacency parameter). This is an attractive
idea since the lexicon is clearly the locus of much cross-linguistic variation: languages have different words.

The outlook at this point was that UG is a rich system of categories and principles, including locality conditions on grammatical relations (with Rizzi’s (1990) Relativized Minimality as a particularly interesting development), X-bar theory, and various ‘licensing conditions’ including conditions on assignment of Case and theta-roles. Discourse-related properties such as topic and focus were also increasingly taken to be elements of grammar, subject to principles and parameters of UG (Horvath 1986; É.Kiss 1987; Rizzi 1997). There were four levels of representation: D-structure, S-structure, PF and LF. The operations were taken to be very simple, possible because the outputs of the operations were subject to strict well-formedness conditions.

The theory was applied and tested on an increasingly wider range of languages (Baker 1988 could be mentioned as a particularly influential work; also Huang 1982; Baker 1996; McCloskey 1996; Cinque 1999; Kayne 2000). It was applied and tested by experimental work on L1 and L2 acquisition. The prediction was that the principles, with their parameters, would manifest themselves relatively early in L1 acquisition. Some of the results appeared to confirm this prediction (see Chien and Wexler 1990; de Villiers, Vainikka and Roeper 1990; Hyams and Wexler 1993; Guasti 2004), although there were always problems (Fodor 2009). It was applied with success to historical linguistics, in conjunction with Lightfoot’s (1979) theory of the role of language acquisition in explaining historical change (see chapters 6, 15, 18, 27, this volume); Roberts (1993), Heycock and Kroch (1993), Kroch and Taylor (2000).
Another theory which is also based on the idea of a rich UG is Optimality Theory
(OT). OT evolved in the early nineties, first as a theory of phonology, but soon also
as a theory of syntax; Prince and Smolensky (1993), Legendre, Grimshaw and Vikner
(2001). OT is a strictly constraint-based framework, where the grammar is seen as
mainly made up of a large set of constraints which are universal but not inviolable.
Instead, linguistic variation is modelled as variation in the ranking of the constraints:
In a language L, a universal constraint C may be ‘ranked higher’ than another
universal constraint C’, and will thereby override constraint C’, in a situation where
there is a conflict between them. In another language L’, C’ may instead be ranked
higher than C, thus overriding C, in a situation where there is a conflict between them.
This will have effects on the output, such that a form which is well-formed in L may
be ill-formed in L’, and vice versa. OT has had a huge impact in phonology
(McCarthy 2001, Prince and Smolensky 2004), less in syntax. There has been a
notable rapprochement between OT and Lexical Functional Grammar, though: see

13.6 UG in the Minimalist programme

The principles and parameters framework with the associated view of UG has
undergone a fairly radical development more recently, starting in the nineties (see
also chapters 7, 31, 32, this volume). One element of this development was
Chomsky’s so called Minimalist Program for linguistic theory articulated by
Chomsky in a series of works starting with Chomsky (1993) and including Chomsky
(1995; 2000; 2001; 2005; 2008; 2013). It is based on the ambition to approach UG
‘from below’, that is to ask why UG has the properties it has, as they have unfolded
in research over the previous decades. One important idea is that some, perhaps
many, of the properties of grammar that have been ascribed to UG are conceptually necessary in any system relating sound to meaning in the fashion of human language, and therefore need not be stipulated as part of UG. The projected outcome of this approach is, therefore, a simpler UG.

The starting point in this approach to grammar and UG is that syntax is a combinatorial system, where the fundamental operation is Merge, an operation which combines two terms, two categories $\alpha$ and $\beta$, and yields a new category $\gamma$. In the strongest, most provocative formulation of the theory Merge is the only operation needed in syntax (Boeckx 2011). Linguistic expressions can be modelled as constructed bottom-up by Merge, deriving LF, a representation interfacing with the conceptual-intentional system, the ‘systems of thought’, and, with the help of morphological and phonological rules, deriving PF, a representation interfacing with articulation and perception. Any other devices, categorial distinctions, conditions on well-formedness, levels of representation, etc. that have been employed in order to formally describe the various languages will have to be evaluated in terms of whether they are conceptually necessary or not, and/or whether they are ‘interface conditions’, not part of syntax proper (‘narrow syntax’), but either semantics/pragmatics or phonology/phonetics. Whatever is conceptually necessary does not need to be stipulated as a property of UG. For example, the representations LF and PF are conceptually necessary, but the representations D-structure and S-structure are not. As such they must either be stipulated as elements of UG, or else eliminated from the

---

9 This is Chomsky’s position. A partly different theory, articulated by Hinzen and Sheehan (2014), is that the syntax is the system of thought, the externalized form of which is the spoken or written language.
theory. In the theory of Chomsky (1993; 1995:ch. 4) they are eliminated, as the conditions previously defined in terms of these levels of representation can be met as they appear in the course of the Merge-based derivation.

The theoretical devices needed within this new model, in its present instantiations, include many of the devices that were important in the earlier model(s): the set of functional categories constructing verbal-sentential projections and nominal projections, movement, including the distinction between A and A-bar movement, theta-roles, case, and agreement. There are some notable differences, though. For example, X-bar theory is no longer included as a stipulation on phrase structure, its properties being instead derived by Merge and labeling, the device which assigns a categorial label to a phrase derived by merge. This is still a source of much debate, though; Hornstein (2009), Sheehan (2013), Chomsky (2013).

A crucially important issue is, as always, how to model linguistic variation, including variation along the diachronic axis. A radical line is that narrow syntax, that is the derivation of LF, is universal. Linguistic variation would then be all within the derivation of PF, the input to the spoken or signed form of a linguistic expression; Berwick and Chomsky (2010), Boeckx (2011), Sigurðsson (2011a,b), Hinzen and Sheehan (2014). The alternative is that syntax allows for at least some variation; Ramchand and Svenonius (2010), Roberts and Holmberg (2010), Holmberg and Roberts (2014).

The idea that Merge is the only operation needed in syntax was helped by a redefinition (first proposed and explored in Chomsky 2004) of syntactic movement as the case when a constituent already merged as part of a tree is merged again. This is ‘internal Merge’, as opposed to ‘external Merge’, when a word or phrase is merged the first time.
Standard versions of the theory, however, include another operation: Agree. This is an operation which can link two terms at a distance, provided the right conditions are met. This theoretical device is dependent on the feature theory also articulated by Chomsky (1993; 1995; 2000) and particularly Chomsky (2001). According to this theory, certain formal features occur in two forms, an interpretable and an uninterpretable form. An example of an interpretable feature is the number feature on a plural DP. An example of an uninterpretable feature is the number feature on a verb agreeing with a plural DP, which is uninterpretable as it does not affect the semantics of the verb but is just an overt marker of the relation holding between the subject and the predicate. Chomsky (2001) proposes that uninterpretable features enter the syntax unvalued, but need to be assigned a value in the course of the derivation. This is the operation Agree.

Locality is partly redefined in terms of phases, where a phase is a syntactic constituent which is in some sense complete, such that syntactic operations, Agree and internal merge, cannot 'see' into it. Finite clauses, verbal predicates, and DPs are phases. Non-local relations, such as between C, at the head of CP, and a wh-phrase inside the sentence would be instances of Agree, subject to locality conditions (relativized minimality, phases), accompanied by movement subject to cross-linguistic variation.

It seems true to say that the most radical pronouncements of the Minimalist Program are not quite matched by shop-floor practices. Arguably the theory or theories currently employed to formally describe and explain observed facts in languages investigated, what most linguists do, are not radically different from the mainstream theory or theories of the pre-minimalist period (no doubt with some exceptions). There are some new devices (unvalued features), some reduction (fewer
levels of representation, no X-bar theory), and there is a more pronounced ambition
to look behind the formal descriptions and explanations, but much of the theory
remains essentially the same. One notable difference, though, is in the conception of
UG and linguistic variation. I will return to this point at the end of the next section.

13.7 The evolution of UG

Another interesting development with consequences for the theory of UG is based on
research and debate on the evolution of language, a field of inquiry which has
attracted a lot of interest in the last two decades.

A notable, highly controversial contribution to this debate is Hauser, Chomsky,
and Fitch (2002). In this paper, a collaboration between a linguist and two biologists,
they try to lay the groundwork for fruitful comparative research on human language
and non-human communication systems. They make a conceptual distinction
between what they call the faculty of language in the broad sense (FLB) and in the
narrow sense (FLN). FLB includes all of the capacities that are relevant for language
and communication regardless whether they specific to language or unique to
humans. A number of elements of FLB are discussed in the paper which are shown
not to be unique to human language, including vocalization, categorial perception,
concept formation, and complex communicative behaviour. FLN would then be the
part of FLB which is unique to humans and specific to language. As they point out
(see also Fitch, Hauser, and Chomsky 2005), it is an empirical question whether FLN
has any content; an alternative idea is that what is unique to humans is a particular,
lucky combination of properties that individually are not unique to humans. They put
forward the hypothesis that a key component of FLN is the capacity for recursion,
making possible the combination of discrete elements (words or morphemes) in infinitely many ways, to form complex meaningful expressions.

How did this capacity evolve? The received view is that what might be called modern human culture, characterised by rapid evolution of technology, use of symbolic objects, and migration and adaptation to new environments is a relatively recent phenomenon, possibly not extending further back than 100,000 years ago (Tattersall 2012, Hauser et. al. 2014). At some point after this time a change took place, in East Africa, which radically altered the conditions for human existence. It is highly plausible that the crucial change was the birth of modern human language. We know that human language based on UG can’t be the result of just gradual evolution by adaptation and natural selection, simply because there hasn’t been enough time. Our closest relatives in the animal world are chimpanzees and bonobos, members of the Pan genus. We know that chimpanzees have a highly limited capacity for concept formation and no capacity for combining concepts in a systematic fashion to form structured expressions (Terrace 1979, 2005, Terrace et. al. 1979). The time since our last common ancestor with chimpanzees is probably between 6 and 7.5 million years ago (Endicott, Ho, and Stringer 2010, Feeney 2014). In this period the capacity for human language must have evolved, and apparently this is not enough time for the evolution of such a specific and complex system to evolve gradually by adaptation and natural selection (Worden 1995, Tattersall 1998).

Taking a broader view, what is minimally needed for human language is the interaction of three distinct components:

(a) a repository of concepts and a capacity for forming new concepts;
(b) a set of operations combining concepts to form structured, complex meanings, particularly forming propositions;\(^{10}\)

(c) a system for assigning physical form to the concepts and the propositions constructed out of them.

Components (a) and (b) are sufficient for thought. Component (c) is necessary to communicate thought. An idea with some plausibility, compatible with the central position of Merge in Chomsky’s minimalist theory, and the comparison of human and non-human language in Hauser, Chomsky, and Fitch (2002), is that the capacity for concept formation evolved separately, and also vocalization and other physical expression of meaning evolved separately. The capacity to combine concepts may also have been present in a rudimentary form, but the big change that happened sometime after the 100,000 year mark was a change which resulted in a chance combination of pre-existing elements, probably ‘rather minor in genetic terms’ (Tattersall 1998), but with hugely advantageous effects, which allowed it to quickly spread in the population. As noted by Hauser, Chomsky and Fitch (2002), it may have been an automatic consequence of increased brain size or it may have been a chance mutation. As suggested in Hauser, Chomsky, and Fitch (2002), and developed further in Berwick and Chomsky (2011) the neural change was the capacity for Merge, making possible the construction of an infinite range of propositions by

\(^{10}\) See Hinzen and Sheehan (2014) on the crucial role of propositions in human language and thought.
combining words, denoting concepts, in the manner of modern human language, thereby radically altering the conditions of human life.\textsuperscript{11}

What does this imply for the idea of UG? The implication of these considerations is that UG, in the strict sense of the innate, genetically encoded capacity for human language, may not be the rich system of syntactic categories, operations, and quite specific principles and conditions (Subjacency, the theta-criterion, the Case Filter, binding principles, the Linear Correspondence Axiom (Kayne 1994), etc.) regulating syntactic operations and defining conditions on representations, with a limited degree of variation permitted by the parameters left open by some of the principles, because consideration of the biological evolution of human language does not allow for evolution of such a rich innate system. This, in turn, implies that the universal properties that have been observed and described in comparative linguistic research may have extra-grammatical explanations more often than used to be acknowledged, at least within mainstream generative linguistics.

13.8 The three factors, linguistic variation, and underspecification of UG
Chomsky (1995) put forward the following model of the design of human language, often referred to as the ‘three factors model’ (see also discussion in §7.2):

15 Factor 1: the genetic endowment, UG.

\textsuperscript{11}See Feeney (2014) for a review of recent research on the birth of language, and a partly different idea of how modern language evolved, with reference to the theory of language and cognition in Burton-Roberts (2011), Burton-Roberts and Poole (2006).
Factor 2: the environment: primary linguistic data for language acquisition.

Factor 3: General principles of computation and cognition.

A possible case of Factor 3 (henceforth F3) would be the principle of economy of derivation which favours derivation (8) over derivation (9) in the example discussed earlier (see also discussion in §24.4). This would plausibly not be anything specific to human language, and probably not even to human cognition, but would be a more general property of animal cognition, or even a more general condition on complex systems in the natural world.

F3 would also include general conditions on learning. An example is what is called Input Generalization in Roberts and Holmberg (2010), Holmberg and Roberts (2014), that is the learning strategy which generalizes from observation of one instance of a category to all instances (cf. also §7.3.2): if you see one white swan, assume all swans are white until further notice. In relation to language acquisition, this might explain the cross-linguistic tendency for word-order generalization across categories, that is the tendency for languages to be either predominantly head-final or head-initial, observed by Greenberg (1963) and discussed in §13.2 above. It would lead the infant learner to generalize from the first instance of a processed head-complement relation to all head-complement relations, with the resulting hypothesis of uniformity across categories subsequently modified in the light of more data. In this perspective, learning a mixed system requires more effort than learning a consistent system, the effort increasing with each counterexample to the first generalization. The prediction is that consistency across categories will be favoured, in a global perspective, which is the case (see Baker 2008). The final output would be
the result of F2 (the environment, the grammar used by speakers in the community) but guided by the Input Generalization strategy, an F3 effect.

F2 would include all those properties which (a) exhibit variation across languages, dialects, and individuals, and (b) are learned by exposure to primary linguistic data. There is a logical connection between (a) and (b): For any grammatical property which can vary, it will have to be learnt which value is instantiated in the language being acquired.¹²

F1, finally, is UG proper, universal properties of human language which cannot be reduced to extralinguistic factors. As before, the way to find out what UG consists of, is to examine putative universal grammatical features and properties, and consider whether they can be explained in terms of extralinguistic factors. If not, they can be assumed to be part of UG.

As mentioned, the view of UG that is implied within this framework, is of a less rich, less specified system, than the view prevalent in classical principles-parameters theory. This has consequences for the approach to linguistic variation. A partly new approach, articulated and discussed in Biberauer and Richards (2006), Holmberg (2010), Roberts and Holmberg (2010), is that linguistic variation can be modelled as an effect of underspecification of UG (see also discussion in §7.3). Adapting the words of Biberauer and Richards (2006), variation occurs where ‘UG doesn’t mind’. In classical principles-parameters theory variation is modelled as

¹² This does not exclude the possibility of properties which are universal among human languages but which are nevertheless learnt; see Biberauer, Holmberg, Roberts, and Sheehan (2014) for an example.
resulting from a set of parameters with specified values, where a language ‘chooses’
a value for each parameter, or, in slightly less abstract terms, the child in the process
of acquiring a language will set each parameter to a value defined by the parameter
on the basis of the primary linguistic data encountered (see chapters 7, 18, 31, this
volume). This is learning. In the case of principles without parameters, no learning is
required. In the current minimalist perspective there are no specified parameters.
Learning is required in every case where no principle or condition, from the F1 or F3
class, determines the form and/or interpretation of a type of expression. This is still
typically – or perhaps even by necessity – a matter of choosing among alternatives
on the basis of the primary linguistic data, but the alternatives are defined by F1 or F3
factors, or a combination of them. In the case of head-complement order there are
only two alternatives, the head must either precede or follow the complement,
because (a) conditions on selection and labelling require merge of a head α and a
complement β (F1), and physical restrictions on articulation and perception dictate
that words must be linearized, so α must either precede or follow β (F3).

As pointed out by Roberts and Holmberg (2010), in the case of variation between
wh-movement and wh-in-situ, formulated in Huang’s (1982) seminal work as the
consequence of a parameter with two specified values, the two alternatives follow if
(a) wh-movement is universal (F1), and (b) movement either precedes or follows
spell out to PF (an F3 effect dictated by the architecture of the grammar). No
specification is required over and above (a) and (b). See Holmberg (2010) for a
demonstration that the classical null-subject parameter of Rizzi (1982) can also
profitably be viewed as a consequence of underdetermination of UG, in conjunction with F1 and F3 (also Roberts 2010; Holmberg and Roberts 2014).\textsuperscript{13}

In a wider perspective there is nothing radically new in the minimalist approach to UG. A generalization, discovered on the basis of comparative research or research on language acquisition, is explained in terms of an innate principle or condition specific to the language faculty, as long as there is no plausible more general principle or other factor which can explain it. This is normal scientific practice. The research agenda may have shifted somewhat, though, within mainstream generative syntactic research, to a more critical attitude towards theoretical devices such as the theta-criterion, the Case Filter, conditions on agreement, anaphoric binding and quantifier-variable relations, standard ingredients of UG ever since Chomsky (1981). The question is if they can be explained in F3 terms, or else can be explained in non-syntactic terms, either semantic or phonological; see Holmberg and Roberts (2014) for some speculations.

\textsuperscript{13} The fact that the order of a given head-complement pair is typically constant in the language of a speaker does not follow from the factors mentioned above. This would be a consequence of the Input Generalization discussed earlier (assume the maximally general hypothesis compatible with the data, a case of F3) and the one form-one meaning principle, which itself is also possibly the effect of a learning strategy: If variation were allowed, the two alternative orders would tend to be associated with different meanings and/or different classes of lexical items. This is, in fact, what is found in languages where head-complement order can vary; see Holmberg (to appear) on VO order in Finnish.
Locality conditions, including the notorious island constraints, are particularly good candidates for explanation in F3 terms, particularly in terms of conditions on human processing capacity which are not specific to language. Proposals to this effect have been made in the literature. I will finish by reviewing some interesting recent research testing whether certain island effects can be explained in extralinguistic terms.

13.9 Another look at constraints on movement
As mentioned above, the constraints on movement first observed and discussed by Ross (1967) make up a particularly well-established set of universal syntactic constraints. For example, movement out of subjects is much more restricted than movement out of objects.

16  a  Which dictator did they make a film about last year?
    b  *Which dictator did a film about appear last year?

Movement out of a finite embedded wh-question is much more restricted than movement out of a finite embedded propositional complement.

17  a  How did you say [ that he fixed the sink <how> ]?
    b  *How did you ask [ whether he fixed the sink <how> ]?

(17a) has a reading where the initial wh-word asks for the manner of fixing the sink (as indicated by the copy within angled brackets). (17b) lacks such a reading, being interpretable only as a question about the manner of asking.
It seems a priori plausible that the judgments could be explained in terms of a constraint on human processing capacity, which may apply more generally than just in the case of grammar. In terms of Chomsky’s three factors, they would be in the F3 class. Proposals to this effect have been made in the literature. Sprouse, Wagers and Phillips (2012) divide these proposals in two categories: One is theories which explain island constraints in terms of independently motivated, non-grammatical constraints on processing capacity. They refer to these as reductionist theories: Givón 1979, Kluender and Kutas (1993), Kluender (1998, 2004), Hofmeister and Sag (2010). The other is theories according to which the constraints are ultimately due to non-grammatical limitations on processing and/or learning capacity, but have been grammaticalized in the course of the cultural evolution of language as grammatical constraints: Fodor (1978; 1983), Berwick and Weinberg (1984), Hawkins (1999) are representatives of this hypothesis. Sprouse, Wagers and Phillips (2012) refer to these as grounded theories of island constraints.

Sprouse, Wagers and Phillips (2012) report a series of experiments to test a prediction made by the reductionist theories, which is that the strength of island effects should vary across speakers as a function of individual differences in processing resources, specifically their working memory capacity. They tested three hundred native speakers of English on a set of island constraints, to see if their judgments of the grammaticality correlated with their performance on two different measures of working-memory capacity. They found no correlation. This is consistent with theories according to which islands constraints are specific to the faculty of language, including theories according to which they derive from UG, as well as with the grounded theories according to which they are grammatical adaptations to constraints on parsing capacity.
In terms of Chomsky’s three factors, classical generative grammar would explain island constraints in terms of UG, i.e. F1. According to the grounded grammatical constraints theory they would be due to F2; they would be acquired by speakers along with other rules of grammar, presumably not as such but as a consequence of acquisition of other properties of structure and movement. According to the reductionist theories they would be F3 effects, being due to non-domain specific constraints on processing capacity. Sprouse, Wagers and Phillips (2012) demonstrate that a particular version of the F3 explanation is untenable. They claim, furthermore, that no other version has been proposed which would be explicit enough to allow testing. See Hofmeister, Staum Casasanto, and Sag (2013) and Sprouse, Wagers and Phillips (2012) for discussion.14

The moral of this section is that claims about extralinguistic explanations of linguistic generalizations, however plausible they may be, especially within the minimalist conception of UG, still need to be subjected to critical examination. And insofar as they turn out to be wanting, the conclusion we are left with may be that the

14 See also Ambridge, Pine, and Lieven (2014), Pérez-Leroux and Kahnemuyipour (2014), Schütze, Sprouse and Caponigro (2014) for a debate of the grammatical or cognitive basis for certain island constraints. Ambridge, Pine, and Lieven (2014) argue that some instances of Subjacency can be explained as effects of a principle of information structure which, they claim, is in principle learnable. In terms of the Chomskyan three factors they would therefore be F2 effects. This is contested by Pérez-Leroux and Kahnemuyipour (2014) and Schütze, Sprouse and Caponigro (2014). The conclusion, if they are right, would be that the island constraints are F1 effects.
generalization is due to UG. The minimalist theory of UG, however attractive it may be conceptually and in the light of considerations of language evolution, is still far from established. Even without specified parameters, UG may turn out to be, after all, an irreducibly rich system of categories and conditions on operations and representations which are not learnt by experience and which have no application outside language, and are therefore part of UG.

References

Ambridge, Pine and Lieven 2014. [PLEASE PROVIDE DETAILS]
Baker, M. 1988. [PLEASE PROVIDE DETAILS]
Berwick, R.C. and Chomsky, N. 2010. [PLEASE PROVIDE DETAILS]
Berwick, R.C. and Chomsky, N. 2011. [PLEASE PROVIDE DETAILS]
Berwick, R.C. and Weinberg 1984. [PLEASE PROVIDE DETAILS]
Biberauer, T. 2011. [PLEASE PROVIDE DETAILS]
Biberauer, T. 2012. [PLEASE PROVIDE DETAILS]
Biberauer, T., Holmberg, A., Roberts, I and Sheehan, M. 2014. [PLEASE PROVIDE DETAILS]
Biberauer, T. and Richards, M. 2006. [PLEASE PROVIDE DETAILS]


Chomsky, N. 1973. [PLEASE PROVIDE DETAILS]


Chomsky, N. and Fitch 2002. [PLEASE PROVIDE DETAILS]


Dahl and Velupillai 2011. [PROVIDE DETAILS]


Dryer, M.S. 2011. [PLEASE PROVIDE DETAILS]

Endicott, Ho and Stringer 2010. [PLEASE PROVIDE DETAILS]

Everett, D. 2005. [PLEASE PROVIDE DETAILS]

Everet, D. 2007. [PLEASE PROVIDE DETAILS]

Feeney 2014. [PLEASE PROVIDE DETAILS]

Fitch, Hauser an Chomsky 2005. [PLEASE PROVIDE DETAILS]

Fodor 1978. [PLEASE PROVIDE DETAILS]
Fodor 1983. [PLEASE PROVIDE DETAILS]

Givón, T. 1979. [PLEASE PROVIDE DETAILS]


Guasti, T. 2004. [PLEASE PROVIDE DETAILS]

Hauser, Chomsky and Fitch 2002. [PLEASE PROVIDE DETAILS]


Hawkins, J.A. 1999. [PLEASE PROVIDE DETAILS]

Hinzen, W. and Sheehan, M. 2014. [PLEASE PROVIDE DETAILS]

Hofmeister and Sag 2010. [PLEASE PROVIDE DETAILS]

Hofmeister, Staum, Casasanto and Sag 2013. [PLEASE PROVIDE DETAILS]

Holmberg, A. 2010. [PLEASE PROVIDE DETAILS]

Holmberg, A. (to appear) [PLEASE PROVIDE DETAILS]

Holmberg, A. and Roberts, I. 2014. [PLEASE PROVIDE DETAILS]


Huang 1982. [PLEASE PROVIDE DETAILS]

Hyams and Wexler 1993. [PLEASE PROVIDE DETAILS]

Kayne. R. 1994. [PLEASE PROVIDE DETAILS]


Kluender 1998. [PLEASE PROVIDE DETAILS]

Kluender 2004. [PLEASE PROVIDE DETAILS]
Kluender and Kuta 1993. [PLEASE PROVIDE DETAILS]


Legendre, Grimshaw & Vikner 2001. [PLEASE PROVIDE FULL DETAILS].


McCarthy 2001. [PLEASE PROVIDE FULL DETAILS]


Pérez-Leroux and Kahnemuyipour 2014. [PLEASE PROVIDE DETAILS]

Pesetsky, Nevis, Rodrigues 2009. [PLEASE PROVIDE DETAILS]

Ramchand and Svenonius 2010. [PLEASE PROVIDE DETAILS]


Roberts, I. 1993. [PLEASE PROVIDE DETAILS]

Roberts, I. and Holmberg, A. 2010. [PLEASE PROVIDE DETAILS]

Rodrigues and Sândalo, to appear. [PLEASE PROVIDE DETAILS]


Schütze, Sprouse and Caponigro 2014. [PLEASE PROVIDE DETAILS]


Sigurðsson 2004. [PLEASE PROVIDE DETAILS]

Sigurðsson 2011a. [PLEASE PROVIDE DETAILS]

Sigurðsson 2011b. [PLEASE PROVIDE DETAILS]

Sprouse, Wagers and Phillips 2012. [PLEASE PROVIDE DETAILS]

Sun et al. 2011. [PLEASE PROVIDE DETAILS]

Tattersall 1998. [PLEASE PROVIDE DETAILS]

Terrace, 1978. [PLEASE PROVIDE DETAILS]


Worden 1995. [PLEASE PROVIDE DETAILS]