There are a number of different distinctions that the labels “competence” and “performance” have been used to highlight. While the associations between the different uses of the terms are clearly not entirely accidental, the in-principle divergences between them are also significant and easy to underestimate, so in this chapter I will focus on separating out the various ways in which these terms have been used. Roughly following Phillips (1996, pp.275–277), I’ll focus on these three:

- The distinction between sentences that are grammatical and sentences that are parsable.
- The distinction between what speakers can do and what speakers choose to do.
- The distinction between static/declarative knowledge and procedural/operational knowledge.

Along with these three, Phillips also discusses the distinction between “what the grammar computes” and “how it does the computing, i.e. what algorithm it uses”. By this he means a certain formal concept, which, as he points out, shouldn’t be taken for granted as a relevant ingredient in theorizing about language. But to the extent that it does turn out to be relevant it will have some close relationship to the third distinction I mention above. For this reason I do not treat it as a fourth independent distinction.

1 Grammatical versus parsable

Given an expression $E$ and a grammar $G$, we can distinguish between (i) whether $E$ is licensed by $G$, and (ii) whether it is possible to work out, using certain bounded computational resources, that $E$ is licensed by $G$.$^1$ To say that $E$ is grammatical is to say that the answer to the former question is yes; this is, of course, relative to a particular grammar. I’ll use the slightly less standard term parsable to refer to the situation where the answer to the latter question is yes. The parsability of an expression is relative to not only a grammar, but also to a particular set of assumed computational limitations.

At the level of principle, this distinction should not be controversial. The substantive question

$^1$For parallelism, the former could be phrased as “whether it is possible to work out, using unbounded computational resources, that $E$ is licensed by $G$”. 
is simply which expressions, if any, are grammatical (relative to some relevant grammar) and yet unparsable given the particular computational resources that the mind brings to bear on language processing.

1.1 The classical case of center-embedding

The classical example that illustrates the distinction between grammaticality and parsability is the case of “center-embedded” sentences. Chomsky and Miller (1963, p.286) propose that while (1) is “surely confusing . . . it is perfectly grammatical”.

(1) The rat the cat the dog chased killed ate the malt

This proposal holds that, although (1) is in fact licensed by the mental grammar of an English speaker, it is not possible to establish this given the computational resources available to the language processing mechanism — in the terminology introduced above, (1) is grammatical but not parsable. This requires, of course, an account of why a sentence like (2) does not lead to an analogous parsing failure (Miller and Isard, 1964).

(2) The dog chased the cat that killed the rat that ate the malt

Miller and Chomsky (1963, p.468) give a formal complexity metric — “degree of self-embedding” — that sets sentences like (1) apart from others like (2). The important point is that (1) is an instance of the unbounded pattern illustrated in (3) below, which Chomsky (1959) had earlier noted cannot be recognized by any finite-memory computational device; whereas (2) is an instance of the pattern in (4), which can be recognized by a finite-memory device. Intuitively, a machine can recognize unboundedly long strings of the sort in (4) by simply repeating ‘that ⟨VERB⟩ the ⟨NOUN⟩’ in a loop without regard to how many repetitions have already passed. Recognizing strings of the sort in (3), on the other hand, requires tracking the number of sentence-initial noun phrases, so that these can be paired up with the subsequent verbs.

(3) a. The rat ate the malt
   b. The rat the cat killed ate the malt
   c. The rat the cat the dog chased killed ate the malt
   d. The rat the cat the dog the girl brought chased killed ate the malt ...

(4) a. The dog chased the cat
   b. The dog chased the cat that killed the rat
   c. The dog chased the cat that killed the rat that ate the malt
   d. The dog chased the cat that killed the rat that ate the malt that pleased the brewer ...

The self-embedding metric is expressed in terms of the structural descriptions assigned to these sentences by phrase-structure grammars, but this can be regarded as a proxy for the deeper point about the capacity of finite-memory devices: the incompatibility of (3) with the finite-memory limitation is what forces the structural descriptions of these strings to exhibit self-embedding (Chomsky, 1959, p.167).
Note that under the (very plausible) assumption that the human language processor is a finite-memory device, there is no way for the unboundedly-many sentences indicated in (3) to all be grammatical and parsable. If we grant, following Miller and Chomsky (1963), that they are all grammatical, then there is no avoiding the conclusion that grammaticality and parsability will “come apart” at some point as we extend the pattern in (3). It is important to emphasize that this conclusion can be reached entirely independently of the observed comprehension differences between (1) and (2). In a parallel universe, it might have turned out that the first ten sentences in (3) were all easily comprehensible, and since that was as far down the list as anyone had looked, it was uncontroversially concluded that the entire unbounded pattern is grammatical. Then, after realizing that finite-memory devices would be unable to handle such sentences past some depth, some researcher made the prediction that at some depth they would, despite their grammaticality, be unparsable; and at that point, someone checked the eleventh sentence in the sequence and the prediction was confirmed.

The fundamental property identified by Miller and Chomsky underlies later accounts of exactly what “fails” on sentences like (1) in terms of explicit incremental parsing algorithms for phrase-structure grammars (Abney and Johnson, 1991; Resnik, 1992). The same ideas can also be carried over to parsing algorithms for simple minimalist grammars (Stabler, 2011; Kobele et al., 2013).

1.2 Logical alternatives

While familiar, this is not the only approach one might take to explaining speakers’ reaction to (1) and how it differs from (2). The other natural possibility is that (1) is simply ungrammatical. Speakers’ judgements of sentences like (1) are not obviously different from judgements of sentences that linguists typically classify as ungrammatical. If Miller and Chomsky’s analysis is correct, then we must accept that we cannot reliably experimentally detect the difference between ungrammatical sentences and grammatical-but-unparsable sentences. The term unacceptable can be used for this shared empirical status, while leaving the question of grammaticality open (Chomsky 1965, pp.10–11; Schütze 2016, ch.2; Leivada and Westergaard 2020).

So what we observe is that (1) is unacceptable; one possible explanation for this is that it is grammatical but unparsable, and another is that it is ungrammatical. The latter certainly seems, at least initially, to be the simpler and more parsimonious alternative, since it allows for a more transparent relationship between observable acceptability facts and underlying grammatical constructs. Upon closer consideration, however, constructing a phrase-structure grammar that treats acceptable sentences like (3a) and (3b) in a linguistically adequate manner while excluding (3c) turns out to be very awkward; for example, whatever mechanism attaches ‘the cat killed’ to ‘the rat’ in (3b) would need to somehow be prevented from attaching ‘the dog chased’ to ‘the cat’ in (3c). Most importantly, a phrase-structure grammar that reigns in the center-embedding pattern in (3) in this way would be no less cumbersome than one that reigned in the pattern in (4). A theorist who concocts the appropriate phrase-structure rules for English to put a bound on the pattern in (3) might therefore expect that, in some other language, we will find a pattern structurally analogous to (4) which will need to be similarly reigned in. But degradations in acceptability and/or com-

---

2This is not strictly correct, because I am glossing over an important technicality: there are certain supersets of the collection of strings in (3) which are recognizable by a finite-memory device (for example, the unconstrained set of all strings). But it can be shown that English sentences do not correspond to one of these supersets (e.g. Partee et al., 1990, p.480).

3“if the grammar \( G \) incorporated in \( M \) exceeds any finite automaton in generative capacity, then we know that \( M \) will not be able to understand all sentences in the manner of \( G' \)" (Miller and Chomsky, 1963, p.467).
prehension with center-embedding structures have been reported across a variety of languages (e.g. Cowper, 1976; Hakuta, 1981; Bach et al., 1986; Hoover, 1992), whereas similar degradations with “right-embedding” patterns like (4) are apparently unattested. In short, analyzing the breakdown in acceptability in (3) as a matter of grammaticality would provide no explanation for why the breakdown is observed in (3) rather than in (4), whereas analyzing it as a matter of parsability explains why the asymmetry runs in the direction that it does; see Pulman (1986, pp.204–206) for valuable discussion emphasizing this point.

When we claim that a grammaticality-based limitation on depth of embedding would fit equally well with center-embedding as with right-embedding, it is important to note that this is a statement about a particular class of grammars, i.e. a particular grammar formalism. The claim, more precisely, is that the particular toolkit provided by the context-free phrase-structure grammar formalism is no better suited to imposing a depth limitation on center-embedding than it is to imposing such a limitation on right-embedding. The argument that the parsability account is more parsimonious than the grammaticality account, therefore, rests on the assumption that this formalism (or something sufficiently similar to it to share this property) identifies the correct range of possible human grammars.4 One way to respond to the parsability account, therefore, would be to put forward a grammar formalism in which depth-bounded center-embedding can be expressed naturally and parsimoniously, but depth-bounded right-embedding cannot. Such a formalism would allow for a plausible analysis of (1) as ungrammatical, and therefore allow us to maintain a (more) transparent relationship between acceptability and grammaticality.

Joshi et al. (2000) emphasize that, while no such option seems to be available for center-embedding, we should not rush to the same conclusion every time we are faced with an embeddable construction where acceptability degrades after a certain depth. They discuss a certain German scrambling pattern which becomes unacceptable beyond depth two. As with (1), the context-free grammar formalism provides no natural way of incorporating a depth limit of two, but Joshi et al. observe that a different formalism, tree-adjoining grammar (Joshi and Schabes, 1997), makes available an analysis of the pattern which naturally predicts that it should not extend past depth two. The conclusion is that the unacceptable scrambling sentences can be naturally analyzed as ungrammatical in the context of the tree-adjoining grammar formalism; whereas without the options made available by this formalism, we would likely have to resort to a grammatical-but- unparsable split of the sort introduced by Miller and Chomsky.

1.3 Beyond center-embedding

The well-known case study of center-embedding sentences provides a useful model for reasoning about other phenomena where there is debate about whether unacceptability arises from ungrammaticality or unparsability. A prominent example that has more recently attracted attention is the case of “island effects”. The predominant view throughout the transformationalist/minimalist tradition (e.g. Ross, 1967; Chomsky, 1973) has been that island effects are cases of ungrammaticality, but an appealing alternative is that the relevant sentences might, like (1), instead simply

4Savitch (1993) formulates a precise version of the central idea (“some finite sets have the property that, if we treat them as infinite sets, then we obtain descriptions that are simpler than the ones we would obtain if we treated them as finite sets”) which makes explicit the dependence on a particular class of grammars. Savitch shows that a pattern like (3) is indeed essentially infinite relative to the class of context-free phrase structure grammars: by the time one constructs enough of the grammar to conform to the pattern up to a certain depth, it would take more grammar-writing work to impose a limit on the depth than it would to leave the depth unbounded. Miller and Chomsky’s discussion implicitly makes the assumption that the same is true of the transformational grammars that they discuss.
be unparsable (e.g. Kluender and Kutas, 1993; Hofmeister and Sag, 2010). Central to this debate is the question of whether independently motivated assumptions about processing mechanisms can predict unparsability in the appropriate cases. Without such independent motivation, appeals to unparsability risk becoming an unconstrained “blank cheque” with little or no explanatory impact; see Sprouse et al. (2012a), Hofmeister et al. (2012) and Sprouse et al. (2012b) for representative discussion.

In addition to unacceptable sentences that are analyzed as grammatical, we find some sentences that might plausibly be described as acceptable and yet ungrammatical. Examples include the “comparative illusion” effect in (5) (Montalbetti, 1984; Wellwood et al., 2018) and the “missing VP” effect in (6) (Gibson and Thomas, 1999).

(5) More people have been to Russia than I have
(6) The patient who the nurse who the clinic had hired met Jack

This suggests a neat two-by-two classification, with run-of-the-mill grammatical sentences (which are acceptable) and run-of-the-mill ungrammatical sentences (which are unacceptable) along one diagonal, and the two “mismatch” scenarios on the other diagonal. This point of view is reasonable as far as it goes, with acceptability understood as a purely descriptive/phenomenological label that apparently happens to cross-cut grammaticality. But it is not at all obvious that “acceptable” and “unacceptable” pick out natural classes in the way that “grammatical” and “ungrammatical” presumably do. In particular, it seems useful to maintain a distinction between acceptability and parsability: the question of parsability, in the sense defined at the beginning of this section, presupposes that there is a grammatical analysis of the sentence to be found, and therefore simply does not arise in cases like (5) and (6) if they are taken to be ungrammatical. Distinguishing acceptability and parsability leaves room for the possibility that, whatever task the relevant processing mechanisms are “succeeding at” in (5) and (6), is not the same as the task that resource limitations lead them to “fail at” in the classic center-embedded examples like (1); this might be particularly relevant given the somewhat unstable semantic interpretations that speakers associate with these sentences, especially (5).

2 What speakers can do versus what speakers choose to do

This is the distinction between the way a speaker’s linguistic knowledge makes some range of linguistic expressions available, and the selections that a speaker makes from that range of options. Chomsky (1995, p.227) likens ignoring this distinction to asking “that a theory of the mechanisms of vision or motor coordination explain why someone chooses to look at a sunset or reach for a banana”.

Fleshing out Chomsky’s analogy a bit further: we can imagine a research community studying the physical make-up of human limbs, investigating the roles and interactions of the various bones, joints, muscles and ligaments in human motion. A natural possibility is that a complete understanding of these mechanisms will provide us with no insight into the fact that, when humans are left to their own devices, certain possible positionings of arms and legs (e.g. reaching for a banana) are unacceptable sentences that are analyzed as grammatical, we find some sentences that might plausibly be described as acceptable and yet ungrammatical. Examples include the “comparative illusion” effect in (5) (Montalbetti, 1984; Wellwood et al., 2018) and the “missing VP” effect in (6) (Gibson and Thomas, 1999).

(5) More people have been to Russia than I have
(6) The patient who the nurse who the clinic had hired met Jack

This suggests a neat two-by-two classification, with run-of-the-mill grammatical sentences (which are acceptable) and run-of-the-mill ungrammatical sentences (which are unacceptable) along one diagonal, and the two “mismatch” scenarios on the other diagonal. This point of view is reasonable as far as it goes, with acceptability understood as a purely descriptive/phenomenological label that apparently happens to cross-cut grammaticality. But it is not at all obvious that “acceptable” and “unacceptable” pick out natural classes in the way that “grammatical” and “ungrammatical” presumably do. In particular, it seems useful to maintain a distinction between acceptability and parsability: the question of parsability, in the sense defined at the beginning of this section, presupposes that there is a grammatical analysis of the sentence to be found, and therefore simply does not arise in cases like (5) and (6) if they are taken to be ungrammatical. Distinguishing acceptability and parsability leaves room for the possibility that, whatever task the relevant processing mechanisms are “succeeding at” in (5) and (6), is not the same as the task that resource limitations lead them to “fail at” in the classic center-embedded examples like (1); this might be particularly relevant given the somewhat unstable semantic interpretations that speakers associate with these sentences, especially (5).

2 What speakers can do versus what speakers choose to do

This is the distinction between the way a speaker’s linguistic knowledge makes some range of linguistic expressions available, and the selections that a speaker makes from that range of options. Chomsky (1995, p.227) likens ignoring this distinction to asking “that a theory of the mechanisms of vision or motor coordination explain why someone chooses to look at a sunset or reach for a banana”.

Fleshing out Chomsky’s analogy a bit further: we can imagine a research community studying the physical make-up of human limbs, investigating the roles and interactions of the various bones, joints, muscles and ligaments in human motion. A natural possibility is that a complete understanding of these mechanisms will provide us with no insight into the fact that, when humans are left to their own devices, certain possible positionings of arms and legs (e.g. reaching for a banana)
are adopted more frequently than others (e.g. touching an elbow to an ankle). The correct theory of the mechanisms might simply say that both of these positionings of limbs, unlike many conceivable others, are within the range of options made available by the biological machinery under investigation, leaving explanations for the choices that humans make, given that range of available options, to come from elsewhere (e.g. facts about bananas and sunsets).

We can adopt a similar division of labour in the investigation of linguistic, rather than physical, abilities. The object of investigation that Chomsky identifies is the collection of underlying (mental) mechanisms that give rise to the range of expressions available to speakers of a language. The proposal is to adopt the working hypothesis, on analogy with limbs, that this can be pursued independently of attempts to explain why speakers choose to say what they say. This could turn out to be somehow misguided, but the analogy highlights the fact that it seems particularly reasonable for physical mechanisms, and need not be any less reasonable for mental mechanisms.

This point is the crux of many recurring debates about the evidentiary role of naturally-occurring language use, in particular in the form of corpora, and about the place in linguistic theory of notions like an expression’s frequency or probability. Under the proposed division of labour, the frequency of a particular expression in a corpus is essentially the frequency with which speakers choose to use that expression, and the status of the expression relative to the linguistic mechanisms that Chomsky identifies as the object of study is just one of the many factors that contribute to this frequency. Contrasting approaches adopt the view that “the primary object of study is the language people actually produce and understand”, in which case it naturally follows that “Language in use is the best evidence we have” (Barlow and Kemmer, 2000, p.xv).6 From this point of view there is nothing to be learned from (the acceptability or unacceptability of) artificially-constructed example sentences that have not been “produced by real people in real discourse situations” (Tomasello, 1998, p.xiii). Artificially-constructed sentences can be useful under the approach that Chomsky advocates, however, in the same way that knowing whether or not it’s possible to touch an elbow to an ankle might be useful for understanding limbs.

Once we adopt this division between linguistic mechanisms and language use, many facts about sentence frequencies or probabilities can be set aside; famously, for example, “surely it is not a matter for the grammar of English that ‘New York’ is more probable than ‘Nevada’ in the context ‘I come from —’ ” (Chomsky, 1964b, p.215). More generally, we should not expect our theories of language to themselves deliver predictions that take the form of a probability distribution over sentences: that is, a single function which assigns a probability to each sentence, such that the sum of the assigned probabilities across all sentences is one, with more frequent sentences receiving higher probabilities. This specific kind of probability distribution is usually what is meant by the terms language model or probabilistic grammar; perhaps unfortunately, these are very general-sounding terms, and this specific kind of probabilistic model has become a particularly prominent way to associate probabilities with entities of linguistic interest. (They are extremely useful for certain purposes and are standard topics in introductory computational linguistics textbooks (e.g. Manning and Schütze, 1999; Jurafsky and Martin, 2000).) So it is worth noting that the entities usually described with these general-sounding terms typically do bundle together all the various factors that contribute to a sentence’s frequency. For example, an “accurate” language model will likely assign probabilities in ways that make both of the following inequalities true, even though

---

6The full sentence from Barlow and Kemmer (2000) is “Language in use is the best evidence we have for determining the nature and specific organization of linguistic systems” (emphasis added), which seems to leave open some questions about exactly how these systems relate to “the language people actually produce and understand”, apparently the primary object of study.
only one of these two sentence-pairs exhibits a difference that is “a matter for the grammar of English”.

(7)  a. \( \Pr(\text{I come from New York}) > \Pr(\text{I come from Nevada}) \)
    b. \( \Pr(\text{I come from New York}) > \Pr(\text{I come New York from}) \)

But not all corpus frequency observations, or probabilistically-stated generalizations, will necessarily conflate “can do” (e.g. (7a)) with “choose to do” (e.g. (7b)) in this way. The issues that arise when we move to some less obvious cases are subtle and not well-understood; my aim in what follows is to lay out some points of reference that might provide helpful framing for the questions that remain open, not to outline a consensus or even completely coherent viewpoint.

The interfering factor in (7a) is the fact that we assume, all else being equal, that it is more common for a speaker to choose to express the thought that they come from New York than for a speaker to choose to express the thought that they come from Nevada. This point itself might be expressed probabilistically as follows, where \( m_{\text{NY}} \) and \( m_{\text{Nevada}} \) are the two chosen meanings/thoughts, whatever they might be exactly. They certainly need not correspond to any particular existing notion of what sentence meanings look like — they are just place-holders for whatever sits at the “pivot point” where all non-linguistic factors are accounted for.

\[
\Pr(\text{I come from New York}) = \frac{\Pr(\text{I come from New York} \mid m_{\text{NY}})}{\Pr(m_{\text{NY}})} \times \Pr(m_{\text{NY}})
\]

\[
\Pr(\text{I come from Nevada}) = \frac{\Pr(\text{I come from Nevada} \mid m_{\text{Nevada}})}{\Pr(m_{\text{Nevada}})} \times \Pr(m_{\text{Nevada}})
\]

Given the assumption that \( \Pr(m_{\text{NY}}) > \Pr(m_{\text{Nevada}}) \) — for reasons that we take to be beyond the scope of our theory\(^7\) — the inequality in (7a) will follow even if the two conditional probabilities shown in boxes in (8) are equal. So we can maintain the idea that there is no linguistic difference between the relationship that ‘I come from New York’ bears to \( m_{\text{NY}} \) and the relationship that ‘I come from Nevada’ bears to \( m_{\text{Nevada}} \).

The normalizing effect of this sort of conditional probability can be seen, in cartoon form, in the relationship between Figure 1 and Figure 2. The columns in Figure 1 represent (fictional) raw frequencies of various sentences of the forms ‘come from X’, ‘hail from X’ and ‘come X from’. The shape of this graph is what a standard language model, in the specific sense introduced above, is designed to match, including irrelevant effects of what thoughts speakers choose to express. The graphs in Figure 2, on the other hand, show conditional probabilities like the two that are boxed in (8), factoring out (at least some of) those irrelevant effects to reveal a clearer picture of the relationships between the three linguistic forms ‘come from X’, ‘hail from X’ and ‘come X from’. This is all assuming that, to the extent that ‘I come New York from’ occurs at all, for example via slips of the tongue, it occurs as an expression of \( m_{\text{NY}} \).

In particular, Figure 2 indicates that, as mentioned above:

\[
\Pr(\text{I come from New York} \mid m_{\text{NY}}) = \Pr(\text{I come from Nevada} \mid m_{\text{Nevada}})
\]

\(^7\)Notice how utterly mysterious a probability like \( \Pr(m_{\text{NY}}) \) remains, despite the pithy notation: it is not the probability that the speaker in fact comes from New York, it is the probability that the speaker chooses to say something that conveys this thought (whether as a truthful statement, or as a lie, or as a joke, etc.). In this simple case it might not be obviously wrong to suppose that \( \Pr(m_{\text{NY}}) \) and \( \Pr(m_{\text{Nevada}}) \) are roughly proportional to the two states’ populations, but things are likely to be significantly more complicated as soon as we even consider, for example, ‘I don’t come from New York/Nevada’.
Figure 1: Raw frequencies

Figure 2: Conditional probabilities
in keeping with the idea that there is no important linguistic difference between these two sentences. The less accidental asymmetry between the two sentences in (7b), on the other hand, persists in Figure 2:

\[ \text{Pr}(\text{I come from New York | } m_{NY}) > \text{Pr}(\text{I come New York from | } m_{NY}) \]

because it was a “minimal pair” to begin with.  

This is one way to restate the widespread view of a relatively uncontroversial case of non-linguistic factors affecting raw frequencies. There is, of course, plenty of room for debate and disagreement, in particular regarding the notion of “meanings” that are conditioned upon: there is (to put it mildly) no agreed upon way to determine whether two sentences are “candidate expressions of a common meaning”, let alone which assumptions analogous to \( \text{Pr}(m_{NY}) > \text{Pr}(m_{Nevada}) \) are justifiable.

A line of work discussed by Bresnan et al. (2001), Manning (2003) and Bresnan et al. (2007) focuses on some cases where teasing apart the effects of “can do” from “choose to do” is trickier. These cases involve the active/passive alternation and the dative PP/double object alternation (‘give X to Y’ vs. ‘give Y X’). But for parallelism we can begin by considering a modified version of our running example: what would we conclude if the conditional probabilities of the form \( \text{Pr}(\text{I come from X | } m_X) \) were not in fact equal for different choices of X? For example, we might imagine that the graphs in Figure 2 had instead come out as in Figure 3. The unexplained observation here would be that the preference for ‘come from’ over ‘hail from’ appears to be stronger with New York and Indiana than it is with Ohio and Nevada. We can identify two possible responses to this kind of scenario.

![Figure 3: Differences in conditional probabilities](image)

The first is to say that there are, in fact, differences between the non-linguistic factors that can lead a speaker to say ‘come from’ and those that can lead a speaker to say ‘hail from’. On this

---

8Keller and Asudeh (2002, p.240) emphasize that asymmetries in raw corpus frequencies will also arise from “performance effects” of the sort discussed in Section 1: for example, we would expect (2) to have a higher frequency than (1), for reasons that are neither analogous to (7a) nor analogous to (7b). In the current section I restrict attention to the difference between (7a)-type asymmetries and (7b)-type asymmetries, but it is oversimplifying to assume that these are the only two issues being conflated in raw corpus frequencies.
view, Figure 3 does not show apples-to-apples comparisons of the appropriate sort; there is no shared $m_{NY}$. If, for example, there is a difference between coming from a place and hailing from a place (whether in truth condition, connotation, information structure, etc.), and Ohioans and Nevadans are more prone to hailing than New Yorkers and Indianans are, then the supposedly puzzling pattern in Figure 3 would be in effect just another instance of the kind of distortion that we see in Figure 1.\footnote{Even before we get to the different patterns in Figure 3, one might also be tempted to appeal to such a difference between ‘come from’ and ‘hail from’ just to explain the consistent asymmetry between these two forms in Figure 2.}

Alternatively, we could suppose that there are no differences between ‘come from’ and ‘hail from’ that non-linguistic factors can tap into, so all the relevant distortions have indeed been factored out in Figure 3, and conclude that there is an interaction that is \textit{linguistic} in nature between the ‘come from’/‘hail from’ alternation and the choice of PP complement. There is nothing peculiar about the idea that linguistic effects might take the form of an interaction like this — it is exactly what we might expect to find if we analyzed corpus frequencies for the four sentences in (9), for example. Taking the alternation between the sentences with existential ‘there’ ((9a) and (9c)) and the sentences without ((9b) and (9d)) to be non-linguistically inert like the ‘come’/‘hail’ alternation, we would expect any skew towards the existential-‘there’ construction to be stronger in the wh-movement sentences than in the non-movement sentences. Figure 4 shows a possible pattern illustrating this effect.

\begin{align*}
(9) \quad & \text{a. What were there pictures of on the wall?} \\
& \text{b. What were pictures of on the wall?} \\
& \text{c. There were pictures of hamsters on the wall} \\
& \text{d. Pictures of hamsters were on the wall}
\end{align*}

\begin{align*}
\Pr(\ldots | m_{\text{what}}) & \quad 1 \\
\text{(9a)} & \quad \text{(9b)} \\
\Pr(\ldots | m_{\text{hamsters}}) & \quad 1 \\
\text{(9c)} & \quad \text{(9d)}
\end{align*}

Figure 4: A fictional pattern of corpus frequencies for the sentences in (9)

The fictional situation illustrated here with Figure 3, along with the two possible interpretations just discussed, is analogous to the situations presented by Bresnan et al. (2001) and Bresnan et al. (2007). Corresponding to the ‘come from’/‘hail from’ alternation in our running example, Bresnan et al. (2001) study the active/passive alternation; they find that an overall skew towards actives over passives is attenuated (not when New York is replaced with Nevada, but) when the agent is third person and the patient is first/second person, for example. In Bresnan et al. (2007), the subject is the dative alternation between ‘give X to Y’ and ‘give Y X’; they find, among other
effects, that the overall skew towards the double object construction (‘give Y X’) is attenuated when the theme argument (X) is animate and/or the recipient argument (Y) is inanimate. What sorts of arguments might bear on the question of whether these effects are linguistic in nature?

There is one view of these findings that suggests an interpretation in terms of non-linguistic differences between the members of the alternating pairs. On this line of thought, there is no shared \( m \) that is common to both ‘The man knows me’ and ‘I am known by the man’, because the surface subject position is associated with topic-itude; therefore the active makes ‘the man’ the topic, whereas the passive makes the speaker the topic. Similarly, there is no shared \( m \) between ‘give them him’ and ‘give him to them’ because the immediately post-verbal position in each case is associated with discourse-old information. If it is a non-linguistic fact, of the \( \text{Pr}(m_{NY}) > \text{Pr}(m_{Nevada}) \) sort, that first/second person speech act participants and other animate referents are more commonly topics or discourse-old information, then the relative prevalence of ‘I am known by the man’ and ‘give him to them’, contrary to the overall skews towards actives and double-object constructions, would follow via the same logic as we saw in (8).

On the other hand, Bresnan et al. (2001) and Manning (2003) emphasize that the relative rarity of actives like ‘The man knows me’ mirrors what is usually described as a categorical ban in some languages, where “the person of the subject argument cannot be lower [on a hierarchy where 1st and 2nd outrank 3rd] than the person of a nonsubject argument. If this would happen in the active, passivization is obligatory; if it would happen in the passive, the active is obligatory” (Bresnan et al., 2001, p.15). They argue on the basis of theoretical parsimony that the effect should therefore be regarded as a grammatical one, not only in these languages where it functions as a “hard constraint”, but also in English where it functions as a “soft constraint”. This would put the interaction between the active/passive alternation and the person of the arguments in the same category as the grammatical effect in (9). One thing that is arguably unusual about this conclusion is that the English effect seems to be less pronounced (as measured either by corpus frequencies or by acceptability judgements) than others that are typically thought to be grammatical in nature, but the existence of these more subtle effects need not be surprising if we accept the idea that there can be degrees of (un)grammaticality (Chomsky, 1964a; Chomsky, 1965, pp.148–153; Keller, 2000).

Deciding between these options is not straightforward, at least given our current levels of understanding. A useful point of comparison, which highlights some of the subtleties involved and relevant gaps in our understanding, is to consider the role that an analogous kind of frequency data has played in the study of phonology. The greater impact of frequency data and probabilistic models in the study of phonology as compared to syntax is likely due, at least in part, to the fact that some of the thorny open questions from the discussion above have more approachable analogs in the case of phonology.

A prominent line of work in phonology engages with frequency data by taking the relationship between an underlying representation (UR) and a surface representation (SR) to be probabilistic: instead of a single licit SR for each UR, there is a probability distribution over possible SRs for each UR. In the various forms of probabilistic Optimality Theory (e.g. Anttila, 1997; Boersma and Hayes, 2001; Goldwater and Johnson, 2003), this amounts to a probability distribution over candidates for each input. The probabilities assigned by these theories are conditional probabilities in the sense illustrated in Figure 2 and Figure 3 above. A data set from Anttila (1997) which appeared frequently in early discussions of these theories concerns alternative realizations for Finnish genitive plurals. The four trisyllabic stems in Figure 5, for example, show differing proportions of the “strong” and “weak” realizations in corpora of naturally-occurring speech (Anttila, 1997, p.58).
Figure 5: Finnish genitive plural realizations

Notice that by taking this sort of pattern as the point of departure, Anttila and others completely bypass distorted raw frequencies of the sort we saw in Figure 1: the issue of how often people choose to talk about cameras or hospitals does not get in the way at all. These conditional probabilities are particularly natural and straightforward in the phonological case, because it is relatively uncontroversial how to classify the observed surface forms into groups of competitors. Taking [kámerojen] and [kámeröiden] to be “non-linguistically equivalent” expressions of a single underlying sense is less precarious than making the same assumption about ‘The dog chased you’ and ‘You were chased by the dog’. The more confident we can be that there are no differences between [kámerojen] and [kámeröiden] for accidents of the non-linguistic world to tap into, the stronger the argument that we should seek linguistic explanations for the differences between the four distributions in Figure 5; the stronger the argument, in other words, that the differences between those four distributions reflect differences in the way linguistic knowledge makes those options available, rather than what speakers choose to do given the available options.

There is, however, another kind of probabilistic model that is widely used in phonology which has a different character, and at least bears a closer resemblance to a case where genuine “speaker choice” proportions of the sort in Figure 1 play an important role; but on closer inspection this is not really the case, for reasons that are illustrative. These are probabilistic models of phonotactics, the constraints on possible word-forms in a language. The data to be accounted for includes the fact that, for example, English speakers reliably rate ‘blick’ as a better possible word than ‘bnick’. Many models of phonotactics take the form of a probability distribution over word-forms, and a model where Pr(blick) > Pr(bnick) is taken to correctly predict the asymmetry in ratings; intuitively, this might arise because ‘blick’, more than ‘bnick’, shares relevant properties with observed words such as ‘blimp’ and ‘lick’. Unlike those in Figure 5, these are not probabilities in a distribution over candidate expressions of some particular sense or meaning — they are “unconditional probabilities”,

12
from a single distribution over all potential word forms, including ‘blick’ and ‘bnick’ and everything else imaginable. All these word forms’ probabilities will sum to one.

Such a model therefore has the formal shape of a language model in the specific sense introduced above, since it defines a probability distribution over all strings. So if the “unconditioned” probability of ‘blick’ under such a model is a good predictor of its well-formedness ratings, one might be tempted to reconsider the possibility that a language model’s distribution over sentences (which attempts to match the shape of Figure 1) stands in a meaningful relationship to well-formedness as well. But there is an importantly different notion of frequency underlying the phonotactic models: these probabilistic models give the best performance when they are trained on “type frequencies”, rather than “token frequencies” (Bybee, 1995; Hayes and Wilson, 2008; Daland et al., 2011). In other words, the frequency with which a certain word is used in naturally-occurring speech or text is not represented in the training data; rather than a corpus in this sense, the training data takes the form of a lexicon. Relevant sorts of frequencies include the frequency of, say, ‘bl’ appearing at the beginning of a lexical entry, but not frequencies of use in any sense. As a result, the probability assigned to ‘blick’ is interpretable as the probability that the next word form added to the lexicon will be ‘blick’, not the probability that ‘blick’ will occur next in some stream of naturally-occurring text or speech.

It is not easy to make sense of the idea of a collection of sentences such that the probability of a sentence appearing in that collection (or as the next sentence drawn from the same source) would be independent of the effects of “New York and Nevada” facts in the way that these phonotactic probabilities are. When a model of sentence probabilities is trained on a collection of naturally-occurring sentences, the resulting probabilities are interpretable as the probability of a certain sentence appearing as the next naturally-occurring sentence, i.e. the next sentence chosen by some speaker being observed. The construction of a sentence, from lexical items according to syntactic rules, seems to be tied to speaker choices in this sense in a way that the construction of a word form, from phonemes according to phonotactic rules, is not.

3 Static/declarative knowledge versus procedural/operational knowledge

This is the distinction between static/atemporal knowledge of the relationships between complex expressions and their subparts, and knowledge of procedures for constructing and analyzing complex expressions (in “real time”) in the ways defined by those atemporally-formulated/stated relationships.

It is not logically necessary that human linguistic systems be organized in a way that reflects this distinction. A prominent example of a position that rejects this distinction is that of Phillips (1996). It is also possible that this division of labour into static and procedural components might turn out to be a convenient strategy for organizing and developing our theories, without corresponding to any joints at which the natural systems being studied are to be carved.

My main goal in this section is not to present arguments for or against making this static/procedural distinction, but rather to clarify what it would mean to make this distinction while taking something along the lines of contemporary minimalist syntax as a theory of the static, atemporal component. On this issue, minimalism specifically can appear to raise complications that were either absent or less significant in earlier versions of transformational grammar, but I want to suggest that this asymmetry between minimalist syntax and other frameworks is only apparent.
The issue hinges on the “bottom up” nature of minimalist derivations, and their relationship to real-time processing. Here is a representative articulation of the concern:

As mentioned above, the MP [Minimalist Program] as a syntactic theory appears to be a step backwards for psycholinguistics (although perhaps not for syntacticians, of course). One of the fundamental problems is that the model derives a tree starting from all the lexical items and working up to the top-most node, which obviously is difficult to reconcile with left-to-right incremental parsing . . .

Ferreira (2005, p.370)

To address this, I’ll begin with an analogy to simple derivational morphology, as it’s often presented in introductory linguistics textbooks (e.g. Fromkin et al., 2000, pp.54–57). In this basic theory, a grammar consists of some collection of words, each with a category, and some collection of affixes, each with a specification its “category-changing effect”. A small grammar of this sort is given in (10), where I’ve used the slash-based notation from categorial grammar (e.g. Steedman, 1996) to indicate the categories of the affixes. The schemas in (11) indicate the interpretation of the slashes.

<table>
<thead>
<tr>
<th>Word</th>
<th>Category</th>
<th>Affix</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>N</td>
<td>re</td>
<td>V/V</td>
</tr>
<tr>
<td>cat</td>
<td>N</td>
<td>un</td>
<td>A/A</td>
</tr>
<tr>
<td>walk</td>
<td>V</td>
<td>able</td>
<td>A/V</td>
</tr>
<tr>
<td>lock</td>
<td>V</td>
<td>er</td>
<td>N/V</td>
</tr>
<tr>
<td>happy</td>
<td>A</td>
<td>ness</td>
<td>N/A</td>
</tr>
<tr>
<td>modern</td>
<td>A</td>
<td>ize</td>
<td>V/A</td>
</tr>
</tbody>
</table>

(11)

\[
\begin{array}{c}
\text{X} \\
\text{Y} \\
\text{X/Y} \\
\end{array}
\quad
\begin{array}{c}
\text{X} \\
\text{Y} \\
\text{X/Y} \\
\end{array}
\]

Some of the morphologically-complex words we can derive from this grammar are shown in (12).

<table>
<thead>
<tr>
<th>N</th>
<th>A</th>
<th>V</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>N/V</td>
<td>A/A</td>
<td>A</td>
</tr>
<tr>
<td>walk</td>
<td>er</td>
<td>un</td>
<td>happy</td>
</tr>
<tr>
<td>V/V</td>
<td>V</td>
<td>V</td>
<td>A/V</td>
</tr>
<tr>
<td>re</td>
<td>V</td>
<td>V</td>
<td>able</td>
</tr>
<tr>
<td>A</td>
<td>V/A</td>
<td>V/V</td>
<td>V</td>
</tr>
<tr>
<td>modern</td>
<td>ize</td>
<td>re</td>
<td>lock</td>
</tr>
</tbody>
</table>

In this system, like in minimalist syntax, a convenient way to reason about the range of possible expressions is to think in terms of small-to-big derivations: derivations that “start with” individual words and affixes and “end with” complex words. For example, to describe the illustrated analysis of ‘relockable’ we can say that we first combine the word ‘lock’ and the affix ‘re’ to form a larger expression with category V, and then combine this with the affix ‘able’; and we can’t attach the affixes in the other order, because after ‘able’ is attached, there’s no longer an expression with
category V for ‘re’ to combine with. In the case of ‘remodernize’, the picture is reversed and the suffix ‘ize’ must attach before the prefix ‘re’.

In this very simple system, it’s relatively obvious that the procedural or temporal terminology need not be interpreted as any more than a convenient way of speaking. It’s clear that the procedural “story” that we tell conveys no more information than is conveyed in a static, atemporal manner by the tree diagrams in (12), i.e. information about the relationships between complex expressions and their subparts. Presenting the theory with the help of the temporal terminology does not tie it to any claims about the temporal order in which things happen when a speaker produces or comprehends a morphologically-complex word. In particular, the right-branching structure of ‘remodernize’, where ‘re’ is introduced last, is not usually interpreted as making claims that would be in conflict with evidence of morpheme-by-morpheme incremental processing (e.g. evidence that speakers are able to predict a morpheme with category V\A after hearing ‘remodern’). 10

Returning to minimalist syntax now: if we consider first syntactic derivations comprising only simple merge steps, then there is no reason to ignore the very same atemporal interpretation of a theory presented via temporal terminology. The fact that ‘John’ is introduced last, and ‘the’ and ‘girl’ are merged first, in the derivation represented by (13), need not be interpreted in a manner that conflicts with evidence of word-by-word incremental processing. The tree in (13) makes an atemporal claim about the relationships between complex expressions and their subparts, and it’s sometimes helpful to give a temporal rendering of that claim. That atemporal claim may or may not be correct, but adding the temporal rendering of the idea does not hinder (or help) the project of supplementing the atemporal claim with theories of the mechanisms that achieve incremental processing (e.g. being able to predict a VP constituent after hearing ‘John will’).

(13)

In a full-fledged minimalist system, involving movement and/or agreement operations as well as merge steps, extracting the atemporal claims from a temporally-described theory becomes less intuitive. A system involving movement operations appears to be inherently procedural, or inherently temporal, in a way that a merge-only system does not. But this deceiving appearance can be overcome by considering the abstract/formal differences between derivations with and without movement, and the way derivations relate to the idea that the properties of certain expressions depend upon the properties of other expressions.

The small-to-big direction of minimalist derivations can be seen as the logical endpoint of the line

10 The logical point here is independent of whether there is any evidence for this degree of incrementality in morphological processing, of course.
of thought that begins with X-bar theory, namely that the properties of phrases depend upon the properties of lexical items. In a merge-only derivation of ‘The girl will buy it’, the properties of the sentence depend upon the properties of the component lexical items. But this dependence is not direct: the information is channelled through certain intermediate expressions, such as ‘the girl’ and ‘will buy it’. The properties of these phrases depend upon their component lexical items, and the properties of the sentence depend in turn on the properties of those phrases.

(14)

In a more complex derivation involving movement, such as a derivation of ‘What will the girl buy?’, it is equally true that the properties of the full sentence (or question) depend on the properties of the lexical items, and that the information is channelled through certain intermediate expressions. In this case, the intermediate expressions include, for example, the structure shown in (15a). One way of identifying what is distinctive about this derivation, involving movement, is that this intermediate expression does play a role in determining the properties of the final derived structure shown in (15b), but is not a subpart of it.11

(15)  

a.

b.

11This point is complicated by the idea that movement leaves copies. But some distinction has to be made between the ‘what’ that is the sister of ‘buy’ in (15a) and the ‘what’ that is the sister of ‘buy’ in (15b) — for example, in terms of unchecked features.
The static tree-structure representation of ‘the girl will buy it’ in (14) graphically represents the atemporal relationships between this full sentence structure and intermediate expressions such as ‘the girl’ and ‘will buy it’, because all the relevant intermediate expressions appear as sub-parts of (14). Our usual notation, however, provides no similar representation of the atemporal relationship that (15a) stands in to (15b). When intermediate expressions are not subparts of the final expression, they “disappear” if all we write on the page is a tree structure like (15b), and this creates a sense that the temporal rendering is indispensible.

What is important for a sentence-processing device is the network of relationships between intermediate expressions, through which information flows from lexical items to full sentences, regardless of whether those intermediate expressions end up being subparts of the final expression in any obvious sense. In merge-only derivations, these two issues are conflated: the graphical layout of (14) atemporally reflects both the relationship between various intermediate expressions and the part-whole structure of the final expression. But (15b) atemporally reflects only the part-whole structure of the final expression, and leaves the information-flow relationships between intermediate expressions with nowhere to live except the “temporal” dimension. The lack of a standard atemporal representation of these relationships between intermediate expressions adds a hurdle that must (and can) be jumped over in order to conceptualize a parsing device’s task in a way that keeps it distinct from “running through a derivation”.

Having noticed this, a notational solution is clear: what we need is a representation that atemporally reflects the relationships between intermediate expressions in a movement derivation, in just the way that a tree like (14) reflects those relationships in a merge-only derivation. This is often known as a derivation tree, but is essentially identical to the notion of the T-marker in early transformational grammar (e.g. Miller and Chomsky, 1963; Chomsky, 1965); see Hunter (2019) [and probably Kobele, this volume] for discussion.

The perspective encapsulated in the derivation tree notation resolves the apparent tension between minimalist derivations and incremental processing not only in principle, but in practice: it is ubiquitous in work on incremental parsing algorithms for formal grammars based on minimalist syntax (Stabler, 2013; Kobele et al., 2013; Graf et al., 2015; Stanojević and Stabler, 2018; Hunter et al., 2019; Berwick and Stabler, 2019). Leaving aside the specific psycholinguistic issues addressed in these works, the broader point that is relevant here is simply that they provide a demonstration of the compatibility of a syntactic theory that is most naturally presented via small-to-large derivations, with incremental parsing.

References


