1. Introduction

On September 11, 1956 – the day that cognitive science was born according to some accounts (Gardner 1985:28), George Miller gave a talk entitled ‘Human memory and the storage of information’ at the Symposium on Information Theory at MIT. Miller’s seminal paper drew attention to the importance of capacity limits in information processing, suggesting that the ceiling might be set at seven chunks of information, plus or minus two. The linguistic evidence for his hypothesis focused on the recall of monosyllabic words, but it was not long before Victor Yngve, a pioneer in the field of computational linguistics, suggested that Miller’s conjecture might also provide a limit on ‘immediate memory’ in sentence production. Indeed, Yngve went so far as to claim that ‘much of the complexity of [English] syntax can be explained on the basis of [this] hypothesis’ (1960:464), as could the direction of language change and the difficulty of certain patterns for child language learners (1998:634).

Miller was not the only person at the 1956 symposium to put forward a seminal idea. Speaking on the same day, Noam Chomsky outlined the need for transformational operations in syntactic analysis (Chomsky 1956). His ideas created a new paradigm for the study of language, proving a foundation for research on the role of working memory in syntactic computation.

At least three separate lines of inquiry have been pursued in studies of how working memory interacts with syntax. The first involves explanations for why certain types of utterances are uninterpretable even though they are the product of well-established grammatical operations. Sentences with more than one level of center-embedding are a classic case (e.g., Yngve 1960:460, Gibson 1998:34ff, Chomsky & Miller 1963:286, Miller & Chomsky 1963:471ff).

[S1 The rat [S2 that the cat [S3 that the dog worried] killed] ate the malt].

A second line of inquiry involves the role of processing in creating preferences for one acceptable pattern over another, as illustrated by the placement of the phrase to the girl in the following two sentences (e.g., Gibson 1998:51ff).

________________________________________________________________________

* I thank John Hawkins for his extensive and insightful observations; I also benefited from comments by Frederick Coolidge, Miho Choo and the editors for their comments.

1 Yngve used this term interchangeably with ‘working memory’ (e.g., his page 464); Miller’s paper uses only the term ‘immediate memory.’
Preferred – short phrase before long phrase:
The young boy gave [PP to the girl] [NP the beautiful green pendant that had been in the jewelry store window for weeks].

Dispreferred – long phrase before short phrase:
The young boy gave [NP the beautiful green pendant that had been in the jewelry store window for weeks] [PP to the girl].

A third set of questions, on which I focus in this chapter, involve a possible role for working memory in explaining various facts that are traditionally attributed to grammatical principles – such as the requirement that a reflexive pronoun have an antecedent in the same clause, or the unacceptability of the strange prohibition against the appearance of the complementizer that in certain types of question patterns.

Mary insists that [Sally underestimates herself]. (herself = Sally, not Mary)

*Who did you say that left early? (compare: Who did you say left early?)

The possibility that grammatical principles might be derived from processing pressures – a highly disruptive notion within linguistics – has been put forward in different forms.

One idea, pioneered by Hawkins (2004, 2014, this volume), proposes that a grammar’s rules and syntactic representations reflect the need to minimize processing cost, including the burden on working memory. Grammatical rules, Hawkins proposes, ‘have incorporated properties that reflect memory limitations and other forms of complexity and efficiency that we observe in performance’ (2014:6).

*Performance-Grammar Correspondence Hypothesis (abridged):
Grammars have conventionalized syntactic structures in proportion to their degree of preference in performance. (2014:3)

A parallel approach, which I have been developing for some time (O’Grady 2005, 2015, 2022), makes an even sharper break from the generative tradition by denying the existence of a grammar at all. Its starting point is the radical line of emergentist thought outlined in Sections 2 and 3, which lay the groundwork for the series of case studies presented in Sections 4, 5 and 6. Section 7 offers some concluding remarks.

2. Linguistic emergentism
Emergentist approaches to language encompass a broad range of inquiry (MacWhinney 2015:9). Nonetheless, they tend to converge on a commitment to the idea that the properties of complex systems, including language, arise from the interaction of simpler and more basic forces and factors. The version of linguistic emergentism that has been dubbed ‘Natural Syntax’ (O’Grady 2022) adopts three further assumptions.
Assumption 1: Direct Mapping:
The mapping between a sentence’s form and its meaning is direct, in the sense that it does not require reference to syntactic structure.

<table>
<thead>
<tr>
<th>FORM</th>
<th>MEANING</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Harry left</em></td>
<td>LEAVE</td>
</tr>
<tr>
<td></td>
<td><em>&lt;h&gt;</em></td>
</tr>
</tbody>
</table>

Figure 1. Direct mapping

This idea contrasts with the more standard view that the relationship between form and meaning is mediated by syntactic representations.

<table>
<thead>
<tr>
<th>FORM</th>
<th>SYNTACTIC STRUCTURE</th>
<th>MEANING</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Harry left</em></td>
<td></td>
<td>LEAVE</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>&lt;h&gt;</em></td>
</tr>
</tbody>
</table>

Figure 2. Mediated mapping

Assumption 2: Algorithmic Orientation:
The mapping between form and meaning is regulated by algorithms that operate in real time in the course of speaking and understanding. Here are two very simple examples that are part of the processing routine needed to manage intransitive sentences such as *Harry left*.

The First-Argument Algorithm:
Map the referent of the nominal onto the first argument position in the semantic representation.

*Harry*  $\Rightarrow$  PRED  
          $<h...>$

The Predicate Algorithm:
Map the event denoted by the verb onto the predicate position.

*Harry left.*  $\Rightarrow$  LEAVE  
                 $<h>$

On this view, there are no grammatical rules per se, even for word order. English is a subject–verb language because of the order in which the two processing algorithms apply.
If they were activated in the reverse order, English would manifest a verb–subject template like Irish, Hawaiian and various other languages.

Assumption 3: Processing Determinism:
A further key claim of natural syntax is that the algorithms involved in mapping form onto meaning (and vice versa) are largely shaped by forces that contribute to processing efficiency (a view also adopted by Hawkins).

… the [processor] should operate in the most efficient manner possible, promptly resolving dependencies so that they do not have to be held any longer than necessary. This is a standard assumption in work on processing, where it is universally recognized that sentences are built in real time under conditions that favor quickness. (O’Grady 2005:7)

Speed in communicating the intended message from speaker to hearer and minimal processing effort in doing so are the two driving forces of efficiency… (Hawkins 2014:48; see also Hawkins 2004:9)

Taken together, the three assumptions that I have just summarized ensure that the mapping between form and meaning – the essential activity of language – is shaped and constrained by processing considerations that have a natural and well-established role in cognition. In this, they differ from the formal laws of grammar on which most analytic work in contemporary syntax is based. I will focus in this chapter on the effects of processing determinism and on its close association with working memory.

3. Working memory
A striking feature of research on working memory and language is the stark divide between studies of working memory as a mechanism and studies of language as a rule-driven phenomenon. Responsibility for the divide is shared by those on both sides.

There is a large and rich literature on working memory, dating back at least to the 1950s. As the other chapters in this volume illustrate, very important issues have yet to be resolved: What are the components of working memory? Are there specialized buffers, such as a phonological loop? Is memory for items distinct from memory for serial order? What is the link between long-term memory and verbal working memory? And so on.

These are important questions, whose investigation is closely associated with linguistic phenomena ranging from the recall of nonce words to the difficulty of comprehending, say, direct object relative clauses compared to subject relative clauses. Crucially, however, the pursuit of these matters pays little attention to the issues that preoccupy linguists. One finds few references in the literature on working memory to, say, Archangeli & Pulleyblank’s 2015 theory of phonological contrasts or Kayne’s (1994) highly influential analysis of relative clauses.

In contrast, work within mainstream linguistics tends to flow in the opposite direction, making quite minimal assumptions about working memory as it explores the intricacies of
sentence processing. This approach equates working memory with ‘the set of cognitive processes involved in the temporary storage and manipulation of information’ (Gathercole 2008:33), a commonly held view in the linguistic literature (e.g., Gibson 1998:2, Hawkins 2004:12, O’Grady 2005:6, Lewis, Vasishth & Van Dyke 2006:447, Jackendoff 2007:13, Schwering & MacDonald 2020:2). On this conception, which I will also adopt here, working memory capacity is closely aligned with a calculus of processing cost built on the following assumptions.

- The less information that has to be processed, the better.
- An operation that can apply immediately is less costly than an operation of the same type that has to be delayed.
- An operation that requires storage of material is more costly than an operation of the same type that does not require that storage.

Consistent with the conceptual divide in the field, my goal is not to offer new insights into working memory, any more than research on working memory is intended to offer new analyses of phonological contrasts or relative clauses. Rather, the point is to explore the possibility that even the most basic properties of working memory, however they are ultimately integrated into a theoretical model, can contribute to a deeper understanding of the workings of human language.

As I will attempt to show in this chapter, this approach has considerable promise, yielding fundamental insights into a series of problems in the syntax of natural language which have traditionally been attributed to highly abstract formal principles of grammar, which now become unnecessary. This in turn opens the door for an emergentist theory of language built around a spare but uncontroversial conception of working memory.

The way forward for this type of research involves a focus on phenomena whose morphosyntactic properties are both well understood and complex enough to strain the mind’s processing resources. Of the many phenomena that fit this description, I will focus here on just three – a restriction on the interpretation of reflexive pronouns, a curious prohibition on phonological contraction in a type of wh question, and a baffling constraint commonly known as the ‘that-trace effect.’ In each case, I will outline the facts and their relevance as informally as possible, avoiding technical details where it is practical to do so. Readers who are interested in a more formal analysis may wish to consult the detailed treatment of these phenomena outlined in O’Grady (2022).

4. A restriction on pronoun interpretation

It has long been understood that prototypical reflexive pronouns in languages such as English seek out a co-argument antecedent (Jespersen 1933:111). Thus, the reflexive pronoun that serves as the second argument of cut in the example below can refer only to Richard, the first argument of the same verb. It cannot refer to Marvin, let alone to some unnamed party.
Marvin just found out what happened. **Richard** cut **himself** while playing with scissors.

Why should this be? An important clue comes from the following real-time scenario for interpreting the reflexive pronoun in the example above.

a. The nominal **Richard** is encountered and is assigned a referent (represented here as the index $r$).
   
   Richard $r$

b. The transitive verb **cut** is encountered and its two-place predicate-argument structure is projected, with **Richard** as the first argument.
   
   Richard cut
   
   CUT $<r\_>$

c. The reflexive pronoun is encountered and identified as the verb’s second argument (represented by the symbol $x$), thereby triggering the search for an antecedent.
   
   Richard cut **himself**.
   
   CUT $<r\ x>$

d. The processor interprets the reflexive pronoun immediately and locally with the help of the previously identified referent of the verb’s first argument, **Richard**.
   
   Richard cut **himself**.
   
   CUT $<r\ x>$

   $\xrightarrow{\_\_ r}$

The first three steps map the words in the sentence directly onto a corresponding semantic representation, without the mediation of syntactic structure or grammatical rules. In the fourth and final step, the just-encountered reflexive pronoun receives an interpretation that links it to the verbal predicate’s first argument.

This way of proceeding has a quite obvious natural motivation: it has the fortuitous effect of resolving referential dependencies immediately and locally, in response to internal processing pressures. At the point where the reflexive pronoun is encountered and identified as the verb’s second argument, only the first argument is immediately available.

---

2. The relevance of locality to processing cost requires careful consideration, especially if it is measured in terms of linear distance, whose relevance to processing cost has been called into question (Lewis, Vasishth & Van Dyke 2006:451). This critique does not apply to locality that is defined in terms of argument structure, as in my proposal.
to serve as antecedent – the correct result, as predicted by the claim that processing pressures shape this particular part of language.

Interestingly, the interpretive facts are mastered very quickly in the course of first language acquisition, despite a paucity of relevant data. (In the Brown corpus, for instance, there are just five third-person reflexive pronouns in caregiver speech directed to Sarah; see https://sla.talkbank.org/TBB/childes/Eng-NA/Brown/Sarah). This makes perfect sense since, in a way, there is nothing for children to acquire; they have only to surrender to the natural impulse to minimize processing cost. The consequence of that impulse is the immediate resolution of the referential dependency by selecting the nearest possible antecedent – a prior co-argument. In other words, ‘learning’ in this instance simply consists of doing as little as possible.

5. A prohibition on contraction
A signature property of English and many other languages is the placement of \textit{wh} words at the left edge of a question, creating a dependency with a sentence-internal lexical item – usually a verb, as in the following example. \( t \)

\begin{center}
\textbf{Who} did Mary see \textit{?} \\
\end{center}

As illustrated here, the dependency is traditionally represented with the help of a gap (\_). Although I will follow tradition in employing the term ‘filler-gap dependency’ alongside ‘\textit{wh} dependency,’ I do not equate the gap with an actual structural position. Rather, I treat it simply as the point in the course of processing at which the dependency can be resolved – a view that is also widely held in the psycholinguistic literature (e.g., Traxler & Pickering 1996, and Pickering 2000; see Hawkins 2004:171-72 for discussion).

A good deal of research has converged on the finding that filler-gap dependencies place significant burden on the resources of working memory.

There is good evidence that \textit{wh} dependencies increase the difficulty of a sentence. Sentences with \textit{wh} dependencies are rated as harder and less acceptable than sentences without \textit{wh} dependencies ... (Phillips 2013a:90; see also Goodall 2004:102 and Hawkins 2004:173, among many others).

This observation brings us to the question that lies at the heart of this chapter: how does processing cost shape the syntax of \textit{wh} dependencies?

\textbf{Want-to contraction}
A very striking first example can be discerned in the effect that a sentence-initial \textit{wh} word can have on the pronunciation of \textit{want to} in sentences such as the following.
Yes-no question; contraction is natural:
Do you want to stay?
\[\downarrow\]
wanna

Wh question; contraction is unnatural:
Who do you want to stay?
\[\downarrow\]
*?wanna

It is a remarkable fact about both language and linguistics that a single casually pronounced word might shed light on the workings of human cognition, but that does seem to be the case.\(^3\) As Getz notes in her review of the literature on the subject (2019:120), the wanna phenomenon has played a ‘momentous role … in shaping modern linguistic theory.’

Two different subtypes of want are in play here, above and beyond the type exemplified in patterns such as I want that bicycle. The first involves just an infinitival complement, as in the following example.

I want to-stay.
\[\uparrow\]

The second involves both a nominal complement and an infinitival complement.

I want Harry to-stay.
\[\uparrow\] \[\uparrow\]

It is the latter pattern that resists contraction when its second argument is a sentence-initial wh word – as in Who do you want to stay? The question is why.

The search for an explanation has yielded a number of spectacular proposals. Perhaps the most intriguing of these, put forward by Jaeggli (1980:242), is that contraction is blocked by an invisible empty category – the case-marked ‘trace’ (\(t\)) left by the operation that supposedly moves the wh word from its initial position between want and to to the left edge of the sentence (in theories of generative grammar).

\(^3\) In fact, the phenomenon appears to extend well beyond the verb want. A similar contrast seems to occur in patterns such as the following, in which the contracted form expectənə seems fully natural only in the first sentence.

Were you expecting to stay?
Who were you expecting to stay?
A case-marked trace blocks contraction. 
(Jaeggli 1980:242; see also Chomsky 1980:160.)

Who do you want [NP to] stay?
↑________________|

A different way to approach the contractibility asymmetry is to consider it from a processing perspective, with attention to two overlapping demands. A first demand involves articulatory processing: contraction is most natural in rapid speech.

**Natural Contraction**

Contraction of the string XY is most natural when X adjoins to Y without delay. (O’Grady 2005:139ff)

It is for this reason that one can say *I want to stay* in a slow and deliberate manner with a pause after each word, including *want*, but that there is no natural pronunciation of *I wanna stay* in which there is a pause after *wan*.

A second demand that shapes the syntax of contraction can be traced very directly to working memory. A long tradition of psycholinguistic research has established that the *wh* filler must be associated with an appropriate verb at the first opportunity. An early statement of this requirement was formulated by Clifton & Frazier (1989: 292 & 297); see also Gibson (1998:54), Aoshima, Phillips & Weinberg (2002:2), Hawkins (2004:174), Wagers & Phillips (2009:396-97), among many others.

*The Active Filler Hypothesis* (paraphrased)

Associate the filler with a ‘gap’ as quickly as possible.

In sum, we have a situation defined by two demands:

i. Phonological considerations call for the immediate adjunction of *want* and *to* if there is to be contraction.

ii. Working memory pressures favor resolution of the burdensome filler-gap dependency at the first opportunity, which arises upon encountering the verb *want*. (Recall that the gap represents the point in time at which the *wh* dependency is resolved, not a structural position.)

Who do you want ____ to stay?
↑

The filler-gap dependency is resolved here.

(cf. *you want who to stay*)
As can be seen here, the two requirements are mutually incompatible in the forbidden pattern: prompt resolution of the \textit{wh} dependency at \textit{want} undermines the opportunity for immediate adjunction with \textit{to} – a prerequisite for contraction.

There is striking independent evidence for this proposal: in patterns where contraction is impeded, \textit{want} has a lengthier articulation and is followed by a prosodic break, suggesting activity at the verb consistent with resolution of the \textit{wh} dependency (Warren, Speer & Schafer 2003).

\begin{quote}
Who do you \textit{want} to stay?
\begin{itemize}
\item \textit{Lengthening and prosodic adjustment at the point where the filler-gap dependency is resolved}
\end{itemize}
\end{quote}

Crucially, there is no such effect in the pattern that permits contraction, since there is no filler-gap dependency to resolve.\footnote{At least one other computational operation is required here, namely the identification of \textit{you} as the first argument of \textit{stay}. Crucially, however, it takes place at a later point (at the verb \textit{stay}) and therefore does not interfere with contraction.}

\begin{quote}
Do you want to stay?
\begin{itemize}
\item \textit{wanna}
\end{itemize}
\end{quote}

In sum, the constraint on \textit{want to} contraction is real, but it does not call for a syntactic explanation involving case-marked traces or any of the other devices in the traditional armory of generative grammar. Rather, it is an emergent property of working memory, which demands quick resolution of \textit{wh} dependencies at the expense of opportunities for contraction.

\section*{6. A baffling constraint on filler-gap dependencies}

Let us turn now to an even more challenging case involving the syntax of \textit{wh} dependencies. Because this phenomenon requires consideration of somewhat more complex syntactic patterns as well as data from languages other than English, I have deliberately left it for the final part of this chapter.

\subsection*{Cross-clausal \textit{wh} dependencies}

I take as my starting point the fact that some languages (e.g., English, but not Russian) allow filler-gap dependencies to extend across a clause boundary.\footnote{In fact, technically, there are no clauses per se in the theory I adopt. The notion ‘clause’ is really a proxy for a predicate-argument structure in a sentence’s semantic representation.}

\begin{verbatim}
4 At least one other computational operation is required here, namely the identification of you as the first argument of stay. Crucially, however, it takes place at a later point (at the verb stay) and therefore does not interfere with contraction.
5 In fact, technically, there are no clauses per se in the theory I adopt. The notion ‘clause’ is really a proxy for a predicate-argument structure in a sentence’s semantic representation.

PRED
<_ ...>
English:
Who did Mary say [that they saw _]?  

Russian (Dyakonova 2009:215):
*Kogo Olga skazala [čto oni videli _]?  

who.acc Olga.nom say.pst.fem that they.nom see.pst.pl

It is not surprising that some languages reject cross-clausal dependencies.  

If the gap is located in an embedded clause, the filler must be transported across the embedded clause boundary… Carrying a filler across a clause boundary results in additional processing cost. (Kluender 1998:253; see also Kluender & Kutas 1993.)

Psycholinguistic research shows … that processing clause boundaries generally lowers acceptability ratings and causes an increase in processing time. (Hofmeister & Sag 2010:383; see also Alexopoulou & Keller 2007:133 & 136.)

There is no reason why all languages should allow cross-clausal dependencies, any more than they should all permit articulatorily difficult consonants – say, uvular ejectives or velar implosives. This is simply one of the points where the linguistic capacities of normal human beings permit variation. The difference between English and Russian is therefore not surprising.

Moreover, from a processing perspective, the more interesting question is not why Russian prohibits cross-clausal *wh* dependencies; it is how English manages to permit them, given their additional processing cost. The key moment for the processor arises at the point where it has to make the transition to the second cause. In a sentence such as *Who did Mary say they met?*, that moment arises at the verb *say*.

Who did Mary *say* …

At this point, the filler-gap dependency is still unresolved. However, the lexical properties of *say* permit a second clause and, therefore, a new set of possibilities – provided that the *wh* dependency can somehow be transferred into that clause. There is good reason to think that this is achieved by reactivating the *wh* word at the onset of the new clause, as illustrated below. (The horizontal line represents maintenance of the *wh* dependency in working memory, not movement.)
Who did Mary say they …

________________________ [who]

↑

Reactivation of the filler at the onset of the second clause

Evidence from a variety of sources supports the reactivation hypothesis.

i. ERP evidence

Left Anterior Negativity, a sign that a wh dependency is being held in memory, is observed at the beginning of the second clause (Kluender & Kutas 1993:608-10; see also Phillips, Kanzina & Abada 2005:423).

What do you think [that the children bought _]?

↑

LAN continues into this region.

ii. Acquisition evidence

When young children produce wh questions containing a cross-clausal filler-gap dependency, they sometimes repeat the sentence-initial wh word at the beginning of the second clause.

What do you think [what pigs eat]?

Who did he say [who is in the box]?

As observed by Lutken, Legendre & Omaki (2020:37), children appear to be reproducing the sentence-initial wh phrase in order to ‘strengthen their memory representation of the filler-gap dependency.’ Similar suggestions have been put forward by Crain, Goro & Thornton (2006:33), Thornton (1990) and McDaniel, Chiu & Maxfield (1995).

iii. Typological evidence

Some languages place a copy of the wh word at the beginning of the second clause.

Romany (McDaniel 1989:569n):

Kas o Demiri mislinol [kas Arifa dikhlā _]?

whom does Demiri think whom Arifa saw

‘Who does Demiri think that Arifa saw _?’

Passamaquoddy (Bruening 2006:28):

Wen Mali wewitahamacil [wen kisiniskamuk _]?

who Mary remember who I.dance.with

‘Who does Mary remember I danced with _?’

So far, so good – but the syntax of wh dependencies raises many additional puzzles.
A costly complementizer

One of the great enduring mysteries of natural language involves the unacceptability of sentences like the one below, in which a filler-gap dependency extends into the subject position of an embedded clause that begins with the ‘complementizer’ *that*.

Subject gap with a complementizer:

*Who do you think [that _ is here]?*

Subject gap with no complementizer:

Who do you think [ _ is here]?

I will call this contrast, first noted by Perlmutter (1968), the ‘*that*-effect.’ (In generative grammar, it is generally referred as the ‘*that*-trace effect.’)

The *That*-Effect

The complementizer *that* is incompatible with a subject gap.

Why should such an effect exist? The most common explanations make reference to one or another principle of Universal Grammar. (An early favorite was the Empty Category Principle, which required that empty positions be ‘properly governed’ by meeting a series of intricate structural conditions; see Chomsky 1986:10ff for details.)

At first glance, it seems unlikely that a phenomenon of this type could be reduced to a working memory effect. After all, *that* is just one small word, with a barely detectable semantics. Nonetheless, upon closer inspection, it is possible to discern a link to processing cost.

A good place to start is with the question of why the absence of the complementizer has the ameliorating effect that it does. Given the general orientation of Natural Syntax, it makes sense to ask whether the complementizer-free pattern might offer a less costly mapping between form and meaning than its counterpart with *that*.

To see how this might work, let’s pick up the acceptable sentence in midstream, right after the matrix verb *think*.

Who do you think …

In the next step, the processor encounters the verb *is*, the nucleus of the second clause.
Who do you think is

The processor encounters the verb in the embedded clause.

The availability of a subject-less verb at the very point where the new clause begins opens the door to the immediate resolution of the *wh* dependency – arguably obviating the need for reactivation.

Who do you think is

Resolution of the *wh* dependency
by associating it directly with the verb *is*

A key piece of evidence for this scenario comes from the fact that the verb in the lower clause is able to fuse with its counterpart in the higher clause.

Who do you think’s here?
(cf. Who do you think _ is here?)

Contraction would not be possible here if reactivation had occurred between the two verbs. As noted in section 5, phonological reduction can occur only if the second element adjoins to the first without delay. The acceptability of the sentence is therefore strong evidence that the *wh* dependency was resolved immediately – before reactivation even became an option.

Compare this scenario with what happens when the embedded clause is introduced by a complementizer.

*Who do you think [that _ is here]?

Under these circumstances, resolution of the filler-gap dependency is substantially more costly. To see this, let’s once again pick up the sentence just after the complementizer makes its appearance.

Who do you think that …

Upon encountering *that*, the processor finds itself committed to two operations that were not required in the case of the complementizer-free pattern.

First, the complementizer must be assigned an interpretation. This essentially involves identifying its function, which is to signal the presence of a clausal argument.
Projection of a second clause after encountering *that*:

Who do you think *that* \\
↑

The presence of *that* signals an embedded clause.

Second, because the *wh* dependency cannot be resolved at this point, it must be reactivated if the search for a gap is to continue.

**Reaction:**

Who do you think *that* \\
[WHO] \\
↑

reactivation of the *wh* filler as the processor prepares to engage with the embedded clause

Upon subsequently encountering the verb *is*, the filler can be associated with the open subject position.

Who do you think **that is** \\
[WHO] ———

resolution of the *wh* dependency by associating it with the verb *is*

This way of proceeding is significantly more costly than the course of action allowed by the complementizer-free pattern, which offers an opportunity to resolve the *wh* dependency without the need for reactivation.

Who do you think is ... \\
↑

The processor encounters an open subject position at the onset of the second clause.

This contrast leads to a promising two-part hypothesis:

- A cross-clausal *wh* dependency requires the least costly computational option.
- Reactivation adds cost that should be avoided if possible.

A precise prediction can now be made.

*The Prediction*

A *that*-effect will occur when the complementizer alone delays immediate access to the verb in the second clause, thereby triggering reactivation.
Some further evidence
This prediction appears to fare well in a variety of different circumstances and languages, offering a straightforward explanation for a series of important facts. I will mention just two such facts here; for more extensive discussion, see O’Grady (2022).

Fact 1: There is no that-effect in direct object wh questions.
The presence of the complementizer should be irrelevant to the acceptability of embedded non-subject questions.

Cross-clausal wh dependency involving a direct object gap:

Who do you think (that) Mary met ?

Even if that was suppressed here, the presence of the subject (Mary) would preclude immediate resolution of the wh dependency at the clause boundary, necessitating reactivation of the wh filler.

Who do you think (that) Mary ...

[who]  
The processor encounters the pre-verbal subject in the second clause, precluding any opportunity for immediate resolution of the wh dependency and triggering reactivation of the wh filler.

Since reactivation is therefore inevitable in any case, the complementizer cannot be held responsible for triggering an operation that could not have been avoided in any case. As predicted, there is therefore no that-effect.

Fact 2: There is no that-effect in languages with post-verbal subjects.
There should be no that-effect in languages such as Italian whose flexible word order delays the opportunity to resolve the wh dependency until mid-clause, where it can ascertained that the subject argument is missing.

Chi hai detto [che ha scritto _ questo libro]? (Rizzi 1982:NN)
who has said that has written ↑ that book
‘Who do you say that wrote that book?’

Because resolution of the wh dependency at the onset of the second clause is therefore ruled out on independent grounds, the presence of the complementizer is benign and there is no that-effect.

In sum, a single word with a sparse semantics can have a very significant impact on processing cost, creating a cascade of effects that include one of the most mysterious phenomena in syntax. The key to this puzzle lies in the realization that the complementizer is not so much the bearer of processing cost as it is the trigger. When that is solely responsible for triggering reactivation of the wh dependency, its presence is toxic. In contrast, its occurrence is benign in cases where reactivation is independently necessitated.
by other factors. No syntactic principle is in play, just the pressure to minimize processing cost where it is possible to do so.

7. Concluding Remarks

In concluding, it is useful to return to Victor Yngve’s attempt to build a processing-based theory of syntax at the dawn of modern cognitive science. The idea was genuinely radical and could well have changed the history of linguistics, had it attained its goal of explaining ‘the utility of much of the observed syntactic complexity in English as an adaptation of grammar to the limited temporary memory available to speakers’ (1998:634). In fact, of course, the particular theory that Yngve put forward (often dubbed the ‘depth hypothesis’) fell short. Miller & Chomsky (1963:474ff) noted one set of problems, and Yngve himself eventually rejected his idea for a different set of reasons (1998:635).\footnote{In contrast, as noted by Cowan (2015), the conjecture by George Miller that inspired Yngve’s idea has aged quite well.}

But dead ends and wrong turns were as commonplace in the early era of syntactic theory as they are today. In Chomsky’s (1956) presentation at the celebrated MIT symposium (and in Syntactic Structures, which was published a year later), he proposed a distinction between two types of sentences: a ‘small possibly finite kernel’ of simple, active declarative sentences derived by applying only obligatory transformations, and a set of sentences derived by applying optional transformations (p. 123). This idea was eventually abandoned, but not the underlying theory, which underwent numerous rounds of revision and reconceptualization – the Standard Theory, the Extended Standard Theory, Government and Binding theory, Principles and Parameters theory, Minimalism… By contrast, there were no second chances for early theories that sought to understand the intricacies of syntax in terms of limitations on working memory.

Now, though, new discoveries and insights have revived interest in that line of inquiry, with somewhat different foundational assumptions but with the same general objective in mind – understanding syntax in terms of processing cost. Why do reflexive pronouns require a local antecedent? Is a case-marked trace really responsible for blocking want to contraction? Why can languages such as Russian and English differ in the admissibility of cross-clausal wh dependencies? How can a mere complementizer trigger the puzzling ripple of syntactic consequences that it does? Could these and other phenomena perhaps be nothing more than emergent properties of working memory?

A further set of challenges stem from the need to bring together currently unaligned streams of research on working memory and syntactic computation. Working memory, in whatever form it exists, is obviously crucial for language-related cognition, but its relevance to actual theories of language use remains ‘scant,’ as Schwering & MacDonald note (2020:11).

it is completely uncontroversial that language comprehension and production processes are constrained by what is commonly called ‘‘verbal working memory capacity’’… and yet the specific mechanisms posited in classic
VWM models are, with only a few exceptions, absent from theorizing about how limited capacities shape language processes. (Schwering & MacDonald 2020:11; see also Caplan & Waters 2013)

This gap in our understanding of cognition must somehow be bridged, an undertaking that will no doubt shed light on both the mechanisms of memory and the computational operations that shapes the properties of language. That is something to look forward to with considerable excitement.
References


Hawkins, John. 2021. Have grammars been shaped by working memory and if so how? This volume.


