1 Introduction

A large body of work in cognitive science is concerned with understanding the constraints on human language comprehension and production. Although questions about language processing fall within the relatively narrow confines of psycholinguistics, there are deep connections between language processing and independently developed research on memory processes within cognitive psychology. This book is about a particular set of computational models of sentence comprehension processes (Engelmann et al., 2020; Lewis and Vasishth, 2005) that seek to explain how one particular conception of working memory constraints comes into play when we comprehend a sentence. The aim is to use an independently developed process model of human information processing (ACT-R) to account for some of the cognitive processes that unfold when a sentence is read or heard. Because of the narrow focus of the work presented here, our discussion of alternative computational modelling approaches will be cursory. This is in no way intended to diminish the importance of these approaches; we feel that approaches such as connectionist modelling and non-linear dynamical systems-based models add great theoretical value to the field and deserve a fuller treatment.

In fact, even within theories that assume working memory constructs in sentence comprehension, this book summarizes only one particular research thread. The work reported here should be seen as a modest contribution towards a broader, longer-term goal: developing competing theories and models of sentence comprehension or parsing processes that can be quantitatively compared against high-quality benchmark empirical data. In an effort to foster reproducibility, and to allow others to use and extend the computational models presented here, all the associated data and code for this book have been made available from the following repository:

https://vasishth.github.io/RetrievalModels

In the next few subsections, we quickly survey some important aspects of previous research on sentence processing, focussing on two important and closely related ideas: the role of working memory constraints and the role of predictive processing.

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

1.1 Working Memory in Theories of Sentence Comprehension

Theoretical and empirical research in sentence comprehension spans a broad range of topics; for comprehensive reviews of the classical theory, see Frazier (1987a) and Pickering and Van Gompel (2006), and for a discussion of some of the recent theoretical developments, see Traxler (2014). Historically, there have been two broad classes of empirical phenomena that have been studied: the effect on comprehension difficulty of complexity (syntactic or semantic/pragmatic) and of ambiguity.

Miller and Chomsky (1963) were among the first to investigate the role of syntactic complexity in sentence processing by indirectly invoking a limit on working memory capacity. They developed a measure of structural complexity that was meant to correlate with memory limitations (Miller and Chomsky, 1963, 480–482): the ratio between the non-terminal and terminal nodes in the tree representation of a sentence (i.e., a global ratio) was taken as a measure of "the amount of computation per input symbol that must be performed" (Miller and Chomsky, 1963, 480). For example, the ratio of non-terminals to terminals in a sentence like *That John failed his exam surprised Mary* is higher than for the extraposed (and easier to process) version *It surprised Mary that John failed his exam*. In related work, Yngve (1960) proposed the depth hypothesis, which stated that the depth of embedding of a phrase was a major predictor of processing complexity. This line of work on complexity continues to be expanded on today.

The double centre-embedding construction is a classic example that illustrates this shift in emphasis from limits on working memory capacity to the constraints imposed on the predictive process. Janet Fodor is cited in Frazier (1985) as noticing that in English, complex multiple centre embeddings are easier to process when the middle verb is missing (i.e., when the sentence is ungrammatical), compared to when the sentence has the correct syntactic structure. Consider the following sentences:

(1) a. *The apartment that the maid who the service had sent over was well decorated.

b. The apartment that the maid who the service had sent over *was cleaning every week* was well decorated.

The middle verb phrase *was cleaning every week* is missing in (1a), rendering the sentence ungrammatical (ungrammaticality is marked with an asterisk, following linguistic convention); compare the ungrammatical sentence with its grammatical counterpart (1b). In an acceptability rating study, Gibson and Thomas (1997) found similar ratings for both sentences, a surprising outcome given that the first sentence is outright ungrammatical. Gibson and colleagues invoked storage overload when holding predictions in memory:

the increased storage cost of holding items in memory is assumed to lead the parsing system to forget a previously generated prediction of an upcoming verb phrase. Gibson's storage cost proposal would predict similar behaviour across languages; however, it seems that German behaves differently from English. In a set of seven reading studies (self-paced reading and eyetracking while reading), Vasishth et al. (2010) found that in English, reading time at the region following the final verb phrase was shorter in the ungrammatical vs. grammatical constructions. This finding from English is consistent with the Gibson and Thomas proposal. However, Vasishth et al. (2010) found the opposite pattern in German: the region after the final verb phrase was read *slower* in the ungrammatical vs. grammatical sentences. This pattern for German has been replicated, and similar patterns were found for Dutch (Frank et al., 2015), but see Bader (2016) for results inconsistent with this claim about German. Vasishth et al. (2010) suggested that German speakers may be able to hold predictions of upcoming verb phrases in memory better than English speakers, because verb phrases in German embedded clauses are always in the last position. German speakers get more exposure to verb-final constructions than English speakers; this is assumed to allow German speakers to maintain predictions for upcoming verbs in German better than English speakers reading English. The explanation is therefore grounded in experience, not some inherent working memory capacity difference between German and English speakers. Incidentally, the ability to maintain predictions seems to be linked to the properties of the language: when German and Dutch speakers who speak fluent English read sentences in English, they no longer show German/Dutch-like behaviour, and behave like English native speakers, reading ungrammatical sentences faster (Frank et al., 2015). This differentiated pattern of responses conditional on the language being currently used suggests that the probabilistic knowledge about syntactic predictions may not transfer across the languages spoken by an individual; if this conclusion turns out to be correct, it would be an interesting avenue of research in bilingualism, where research often presupposes transfer effects across languages.

Some of the other recent empirical work on reading that is concerned with the role of prediction in processing complex syntactic structures is Levy et al. (2012, 2013), Levy and Keller (2013), Vasishth et al. (2018), and Linzen and Jaeger (2016). However, working memory limitations may also play a role independent of the constraints imposed by predictive processing. For example, Safavi et al. (2016) showed that in Persian, readers tended to forget highly predictable particles in verb-particle constructions; this is unexpected under the probabilistic prediction accounts. Further, Husain et al. (2014) found that strong predictions about upcoming materials could override forgetting effects, but weak predictions did not.

Working memory limitations have also been invoked to explain the effect of ambiguity on comprehension ease. For example, Frazier proposed two heuristic principles that guide parsing decisions. These were mainly directed at explaining so-called garden-path sentences, which are characterized by a local ambiguity in the syntactic structure that is resolved later in the sentence, leading to a possible misparse.

The first principle is Minimal Attachment, which stipulates: "Choose the structurally simplest analysis (the one with the fewest additional nodes)." An example is:

(2) The lawyer knew the answer was wrong.

Here, the parser initially assumes incorrectly that *the answer* is the object of *knew*, because this is a simpler structure than the correct one, in which a missing complementizer *that* appears after *knew*:

(3) The lawyer knew that the answer was wrong.

The second heuristic principle was Late Closure: "Integrate current input into current constituent (when possible)." An example sentence is:

(4) After the student moved the chair broke.

Frazier suggested that Minimal Attachment and Late Closure are reflexes of a constrained capacity working memory system. Regarding Late Closure, (Frazier, 1979, p. 39) writes:

> It is a well-attested fact about human memory that the more structured the material to be remembered, the less burden the material will place on immediate memory. Hence, by allowing incoming material to be structured immediately, Late Closure has the effect of reducing the parser's memory load.

Similarly, regarding Minimal Attachment (Frazier, 1979, p. 40) writes:

> [T]he Minimal Attachment strategy not only guarantees minimal structure to be held in memory, but also minimizes rule accessing. Hence, [Minimal Attachment is also an "economical" strategy] in the sense that [it reduces] the computation and memory load of the parser.

The minimal attachment proposal has an interesting twist. Swets et al. (2008) showed that task demands can modulate whether participants engage in any attachment at all. In other words, participants may be engaging in underspecification, and one possible explanation for why they underspecify may have to do with working memory limitations. To understand the phenomenon of underspecification, consider the triplet of sentences shown in (5):

(5) a. Low attachment:
The son of the princess who scratched herself in public was terribly humiliated.

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

b. High attachment:
The son of the princess who scratched himself in public was terribly humiliated.

c. Globally ambiguous:
The maid of the princess who scratched herself in public was terribly humiliated.

Under the classical account, discussed for example in Frazier and Rayner (1982), the parser should find it easier to complete low attachment than high attachment (compare 5a and 5b) to minimize effort as discussed above, and in the globally ambiguous case (5c), the parser should automatically take the route of least effort and make a low attachment. As a consequence, the globally ambiguous condition should show the same processing difficulty as the low attachment condition. Surprisingly, the relative clause in the globally ambiguous condition has been found to be read *faster* than in the low attachment condition (Traxler et al., 1998); this phenomenon is called the ambiguity advantage.

Swets and colleagues suggested that the ambiguity advantage could be due to an underspecification process under different task demands. To show this, they carried out a self-paced reading study, asking participants to read sentences like (5). They asked different kinds of questions about these sentences, changing the complexity and frequency of the questions in a between-participants design. Forty-eight participants were asked questions about relative clause attachment on every experimental trial. An example question for the above set of example sentences is *Did the maid/princess/son scratch in public?* A second group of 48 participants was asked superficial questions. An example for the above sentences is *Was anyone humiliated/proud?* A third group of 48 participants was asked superficial questions only occasionally (once every 12 trials). Swets and colleagues found that an ambiguity advantage was observed when questions were superficial, but no ambiguity advantage was observed when the questions were about the relative clause attachment (here, the globally ambiguous and low attachment conditions patterned together, as the classical theory by Frazier would predict). Thus, when participants do not need to engage deeply with the target sentences, they may engage in more superficial processing, to the extent that they may not even build completely connected syntactic structure. Although the driver of underspecification here is externally imposed task demands, working memory limitations may also be an additional factor. In a Spanish reading study using eyetracking, von der Malsburg and Vasishth (2013) suggested that low working memory capacity participants may underspecify more often in the face of temporary ambiguity; also see Traxler (2007).

1.2 Prediction in Sentence Processing

In his Syntactic Prediction Locality Theory (Gibson, 1998) and his subsequent Dependency Locality Theory, Gibson (2000) formalized the idea that the parser predicts upcoming material, and that there are limits on how much information can be stored. Storage cost has empirical support from several studies; examples are the double centre-embedding work on English by Gibson and Thomas (1997), discussed above, and a Hindi eyetracking corpus study (Husain et al., 2015). A very different perspective on predictive processing was developed through the work of Jurafsky (1996), Hale (2001), and Levy (2008), among others; the assumption here is not that prediction is constrained by working memory limitations, but rather by an underlying probabilistic grammar representation. As a sentence is processed incrementally, an essentially parallel, or ranked parallel set of possible continuations is predicted, and as one transitions from one word to the next, the change in the probability mass of the predicted continuations indexes processing difficulty. Briefly put, rare continuations are hard to process. These prediction-oriented theories represent a distinct class of account that has two characteristics: it has no need for any constraints imposed by working memory, and it only focusses on "forward-looking processes", i.e., predictions about upcoming material. Extreme forms of prediction theories assume, implicitly or explicitly, no limit on the number of proposed continuations (i.e., massively parallel predictions); for discussion, see Boston et al. (2011). Contrast this with the discussion about ambiguity resolution above, where the focus was on the constraints on accessing previously encountered material. For example, when an attachment site for a relative clause is searched for by the parser, the search is directed towards accessing previously processed material. Such "backward-looking processes" could be subject to somewhat different constraints than "forward-looking processes".

Explicit rejections of working-memory based accounts of sentence comprehension difficulty come from the connectionist modelling literature; these can also be seen as a class of prediction-based models. For example, MacDonald and Christiansen (2002) wrote an important critique of Just and Carpenter (1992), who had claimed that high- and low-working memory capacity individuals process sentences differently. Just and Carpenter present data showing that high capacity participants exhibit smaller differences in object vs. subject relative clause difficulty than low-capacity participants.[1] MacDonald and Christiansen argued that the differences in processing difficulty attributed to inherent capacity differences may be due to an interaction between experience

[1] It is worth noting here as an aside that capacity was measured using the Daneman and Carpenter (1980) reading span task, which may index experience with language rather than inherent capacity per se (Wells et al., 2009).

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

with language and biological (neural architectural) factors that have nothing to do with the capacity of a separate working memory system.

1.3 Working Memory and Prediction as Explanations for Processing Difficulty

In summary, memory load and limits on working memory capacity are candidate explanation for certain aspects of language processing, but certain other aspects of processing have a better explanation in terms of probabilistic predictive processes. Much of the inspiration for memory explanations came, either directly or indirectly, from work in cognitive psychology. Decay and similarity-based interference are two key constructs that have been invoked in psycholinguistics in one form or another; these ideas come from research on memory in psychology (Brown, 1958; Keppel and Underwood, 1962; Peterson and Peterson, 1959; Waugh and Norman, 1965). This connection between sentence comprehension difficulty and research on decay and/or interference has been explored in detail by Lewis (1993, 1996), Gibson (2000), and Just and Carpenter (1992). The critical role that prediction plays in human sentence processing was recognized quite early in connection with formal theories of parsing, as discussed in Resnik (1992). The seminal work of Jurafsky (1996) laid the foundations for the use of probabilistic grammatical knowledge in explaining sentence comprehension difficulty; this line of thinking resulted in another important paper by Levy (2008).

1.4 Current Beliefs about Constraints on Sentence Comprehension

Given the above short (and incomplete) survey, some of the broad tentative conclusions that the last 60 years of work on sentence processing can be summarized as follows. Of course, not everyone will agree with this summary; but in our opinion, the claims listed below are relatively well supported by the literature.

(i) The parser builds incremental structural (syntactic) representations during online processing, although the parser may also, under certain circumstances, engage in underspecification of structure or track only local collocational frequencies (Frazier and Rayner, 1982; Swets et al., 2008; Tabor et al., 2004; Traxler et al., 1998).
(ii) The parser probabilistically predicts upcoming material (Hale, 2001; Levy, 2008).
(iii) What is retained in memory and what is predicted during parsing is probably constrained by a working memory component (Gibson, 1998, 2000; Husain et al., 2014, 2015; Safavi et al., 2016).

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

(iv) Experience with language affects our probabilistic knowledge of language, and consequently, our comprehension (MacDonald and Christiansen, 2002; Wells et al., 2009).

1.5 Some Gaps in the Sentence Processing Literature

Although the last 60 years have seen significant advances in our understanding of sentence comprehension processes, we see several major gaps in existing work (there may be others; these are just the ones that stood out for us during the modelling work reported here).

1.5.1 The Relative Scarcity of Computationally Implemented Models

The first gap is that, instead of making computational/mathematical modelling the basis for theory development, the field has largely relied on verbally stated models of comprehension processes. This has led to a great deal of vagueness in theory development. Verbally stated theories have the great advantage that nascent ideas can be quickly sketched out. Indeed, computational models usually begin with an informal statement of the key intuitions. In psycholinguistics, researchers stop too often at the verbal theorizing stage and never attempt to implement their models. There are of course exceptions to this: some examples of implemented models are listed below.

(i) Connectionist models: Engelmann and Vasishth (2009); Frank (2009); Linzen and Leonard (2018); MacDonald and Christiansen (2002); Rabovsky and McRae (2014).
(ii) Constraint-based models: McRae et al. (1998).
(iii) Probabilistic parsing models: Hale (2001); Levy (2008); Rasmussen and Schuler (2017).
(iv) Dynamical systems approaches: Cho et al. (2017); Smith et al. (2018); Tabor et al. (2004); Vosse and Kempen (2000).
(v) Computational cognitive models of underspecification: Logačev and Vasishth (2016).
(vi) Models of decision processes in parsing: Hammerly et al. (2019); Parker (2019).

As an aside, we note that, with some rare exceptions, one major problem with much of the modelling has been the lack of publicly available reproducible code that allows the reader to independently evaluate or extend the published model.

Despite the fact that several serious attempts exist at implementing theories as computational models, many theoretical proposals remain unimplemented. Except for the simplest of ideas, it is generally not sufficient to stop at verbal

statements. This is because informally stated theories usually have hidden degrees of freedom that allow the researcher to explain away or simply ignore counterexamples. Computational implementations force the researcher to confront the distance between theory and data. Although computational models also have hidden degrees of freedom, these are usually easier to see.

An example of the problems that arise in verbally stated theories comes from the Dependency Locality Theory (Gibson, 2000). Originally, a central tenet of the theory was that only new discourse referents can disrupt dependency completion; in previous work, this point was explicitly brought up by showing that a pronoun, which introduces a given or easily inferable referent, causes less disruption than a newly introduced discourse referent (Warren and Gibson, 2005). Moreover, in the classic description of the model (Gibson, 1998) and in its follow-up revision (Gibson, 2000), the following assumption is adopted: "Although processing all words probably causes some integration cost increment, it is hypothesized here that substantial integration cost increments are caused by processing words indicating new discourse structure" (Gibson, 1998, 12). However, in Gibson and Wu (2013), previously introduced (old) discourse referents are assumed to lead to increased dependency completion cost in exactly the same way that new discourse referents do (Hsiao and Gibson, 2003), without any discussion about the change in the assumptions of the model. This change is actually not an inherently important one for the theory, because one could have easily assumed from the outset that all intervening discourse referents (regardless of whether they are new or old) cause processing difficulty. Nevertheless, the example illustrates that model predictions can be "computed" (in the researcher's mind) without noticing that the model assumptions have changed. A further disadvantage of verbally stated theories is that no quantitative predictions can be derived. This affects the kinds of scientific questions one can ask, and the way that one frames one's predictions. With verbally stated theories, we can only ask questions of the type "Does this effect exist or not?" This kind of framing makes it irrelevant whether the effect is 2 ms or 200 ms in magnitude. As discussed in Section 1.5.4, ignoring the magnitude of the expected effect has important consequences for inference. By contrast, a quantitative modelling approach allows us to focus on how the empirical estimates (and the uncertainty associated with these estimates) match up with the range of predictions from the computational models of interest.

A commonly heard objection to computational modelling is that we don't yet know enough about the process of interest to implement it; a related objection is that an implemented model will miss crucial aspects of the cognitive process of interest. These objections are valid, to some extent. But models should be seen as useful lies that help us see the range of possibilities that could constitute truth (Epstein, 2008). As the word itself suggests, a model is rarely intended to accurately capture every single aspect of reality. The criticism that a model fails

to capture this or that detail points to an important limitation of the model, but is not a reason to abandon the entire enterprise of model development (Smaldino, 2017).

In contrast to sentence comprehension research, within other areas adjacent to cognitive science – artificial intelligence and mathematical/cognitive psychology – the development of different computational cognitive architectures and frameworks has flourished and has had a major and positive impact on our understanding of the phenomena under study. This is because it is well-understood in these areas that computational models allow the scientist to build detailed process models of human cognitive processes and to investigate the quantitative predictions arising from these models. Prominent examples from classical artificial intelligence research are the SOAR (Laird, 2012) and ACT-R (Anderson et al., 2004) architectures; in psychology, the E-Z Reader (Reichle et al., 2003, 2009) and SWIFT (Engbert et al., 2005; Rabe et al., 2020; Richter et al., 2006) models of eye-movement control stand out as examples of comprehensive architectural frameworks of a particular cognitive process of interest (reading). Cognitive psychology has a rich tradition of such models: the working memory models by Oberauer and Kliegl (2006), Lewandowsky et al. (2008), the 4CAPS architecture (Just et al., 1999; Just and Varma, 2007; Varma, 2016) (www.ccbi.cmu.edu/4CAPS/), and other models (Busemeyer and Diederich, 2010; Farrell and Lewandowsky, 2018; Lee and Wagenmakers, 2014; Lewis, 2000). By contrast, in the narrower field of sentence processing, not as much effort has gone into developing comprehensive architectures; an interesting exception is CC READER (Just and Carpenter, 1992; Just and Varma, 2002).

1.5.2 *A Focus on Average Behaviour and Neglect of Individual-Level Differences*

The second gap in current work is that the vast majority of the empirical and modelling work has focussed on explaining average behaviour. Researchers have pointed out that the excessive focus on modelling average behaviour is problematic; for example, see the discussion in Kidd et al. (2018).

As Blastland and Spiegelhalter (2014) put it, "The average is an abstraction. The reality is variation." The average response is not sufficiently informative about the true nature of the cognitive process of interest. The focus should be on understanding the causes for average as well as the individual-level behaviour; this will lead to a better understanding of the systematic reasons that lead speakers/comprehenders to show differentiated behaviour. Individual differences have been investigated in some sentence processing studies (e.g., Just and Carpenter, 1992; MacDonald and Christiansen, 2002; Van Dyke et al., 2014), but the field would benefit from making this a routine part of the investigation of the causes of processing difficulty.

1.5.3 *The Absence of High-Precision Studies*

The third gap in the literature is the absence of properly powered experimental studies. The proliferation of underpowered studies has led to a range of invalid inferences in the literature. The term "invalid inference" here doesn't mean the inferences don't reflect the truth, but rather that they are not supported by the statistical analyses. The most egregious example of invalid inferences is concluding that the null hypothesis is true when the p-value is greater than 0.05.

The underlying reasons for the proliferation of underpowered studies is easy to work out. First, researchers are incentivized to publish "big news" papers as fast as possible; this encourages small sample "microstudies" that seem to lead to groundbreaking discoveries. Second, many of the early discoveries in psycholinguistics were large effects that didn't even need an experiment to establish. An example is the strong garden-path sentence *The horse raced past the barn fell*; one can "feel" the oddness of the sentence even without doing an experiment. Another example is the late closure example mentioned above, *After the student moved the chair broke*; one immediately senses that something is wrong with this sentence. Processing difficulties in such easily discernible effects can be reliably detected even with relatively modest sample sizes. For example, the classic garden-path study by Frazier and Rayner (1982) had only 16 subjects, and 16 items for a four-condition late-closure design, and 16 items for a four-condition minimal attachment design (the late closure/minimal attachment manipulation was between items). This sample size might have been sufficient to detect large effects. But such a sample size is certainly too small to investigate predictability (Vasishth et al., 2018) or memory effects (Jäger et al., 2017, 2020; Mertzen et al., 2020a). For investigations of such subtle phenomena in sentence comprehension, there has been no systematic attempt to assess whether sample sizes used for classical garden-path effects would suffice. The consequence has been that a lot of the data published in psycholinguistics is likely to come from underpowered studies. As we discuss in Section 1.5.4, this has very bad consequences for theory development.

1.5.4 *Unclear Desiderata for a Good Model Fit*

In psycholinguistics, there are no general standards on how to quantify a good model fit. Sometimes root mean squared deviation is used; this quantifies the average deviation from the observed value. But there are better approaches. We discuss two important criteria below that we feel are appropriate for modelling work in psycholinguistics; both are related and are fundamentally graphical in nature.

The Roberts and Pashler (2000) Criteria In an influential paper, Roberts and Pashler (2000) pointed out that a quantitative model's fit to the data

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

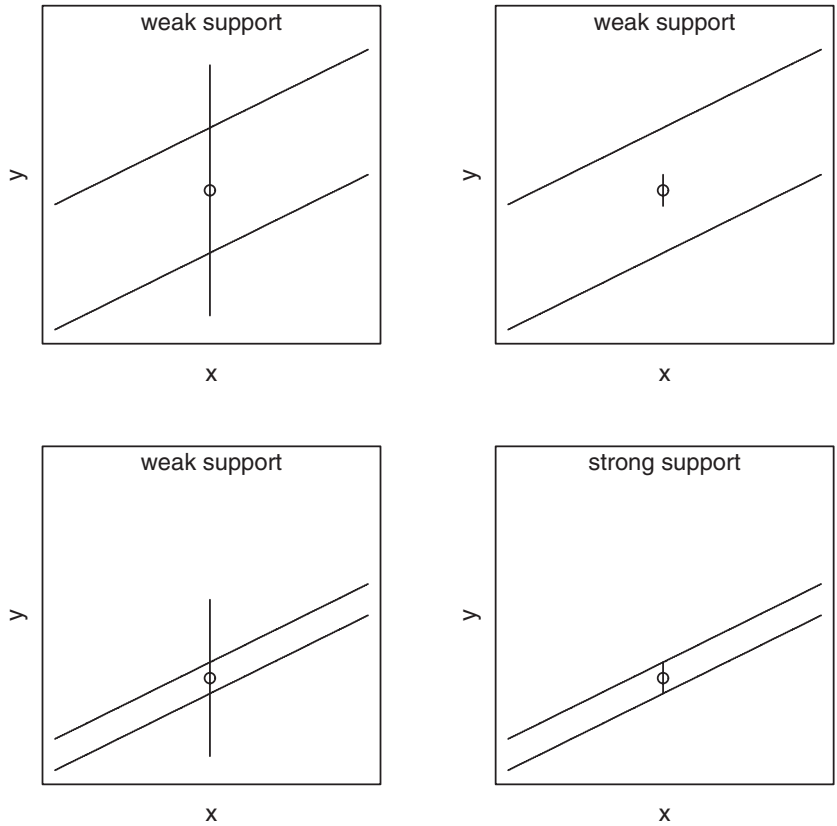

Figure 1.1 A schematic summary of the Roberts and Pashler (2000) discussion regarding what constitutes a good fit of a model to data. The data are represented by the circle (the estimated mean) and the vertical uncertainty interval, and the model predictions by the diagonal parallel lines. If a model predicts a positive correlation between two variables *x* and *y*, strong support for the model can only be argued for if both the data and the model predictions are highly constrained: the model must make predictions over a narrow range, and the data must have low uncertainty associated with it.

is only convincing when two conditions are met. First, the model must make sufficiently constrained predictions. Second, the data should have low uncertainty. These criteria are illustrated in the schematic plot shown in Figure 1.1. The two diagonal lines illustrate a hypothetical range of correlations between two variables *x* and *y* that are predicted by some model. The uncertainty (variability) in the correlation can be high or low. High variability is shown by widely

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

separated lines and amounts to a relatively unconstrained prediction from the model. Roberts and Pashler (2000) make the point that a model's predictions are not going to be impressive if they allow just about any outcome. A more tightly limited prediction will pose a stringent test for the theory. Similarly, a good fit to a model's predictions will be unimpressive and unconvincing if the data have high uncertainty; in practice, what high uncertainty means is that the standard error of the estimated effect is large.

Thus, for a model fit to be convincing, two conditions must be satisfied: the model must make highly constrained predictions and the data must deliver estimates with low uncertainty.

Why Is High Uncertainty Undesirable in the Estimate from the Data? It seems obvious enough that a model should not allow any possible empirical outcome; such a model is not particularly useful because it can "explain" any outcome. Examples from psychology of models that can predict any outcome are discussed in Roberts and Pashler (2000). It is less clear intuitively why empirical estimates of effects need to be measured with precision. As researchers, we are trained to only check whether an effect is statistically significant or not; it is considered irrelevant whether the standard error of the effect is large or not. Here, we show why the precision of the estimate (roughly speaking, the standard error) is a crucial component when evaluating theoretical predictions. The *p*-value is in most cases useless, especially when considered as the sole piece of information from a data analysis (Wasserstein and Lazar, 2016). The limitations of using *p*-values alone for inference is by no means a new insight, but it has been generally ignored in psychology and linguistics.

Estimates of an effect that have high uncertainty (wide standard errors) are also studies that are likely to be underpowered. This has all the bad consequences that come with low power, most dramatically Type M and S errors (Gelman and Carlin, 2014). Type M(agnitude) error refers to an overestimation of the effect magnitude, and Type S(ign) error refers to an incorrect sign (incorrect relative to the predicted or expected effect). Both types of error occur when power is low, as the following simulation demonstrates. Suppose that a true effect in a reading time experiment has magnitude 20 ms, and that standard deviation is 150 ms. In such a situation, a paired *t*-test with a sample size of 26 yields 10% power. If one were to repeatedly run an experiment with this sample size, as shown in the upper part of Figure 1.2, apart from there being many null results, *all* significant results will be either overestimates or will have the wrong sign (or both). By contrast, as shown in the lower part of the figure, when power is high (say, 80% or higher), most significant effects will be close to the true value.

Overestimates or effects with possibly the wrong sign are problematic for the modeller, because the target for modelling itself is misleading.

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

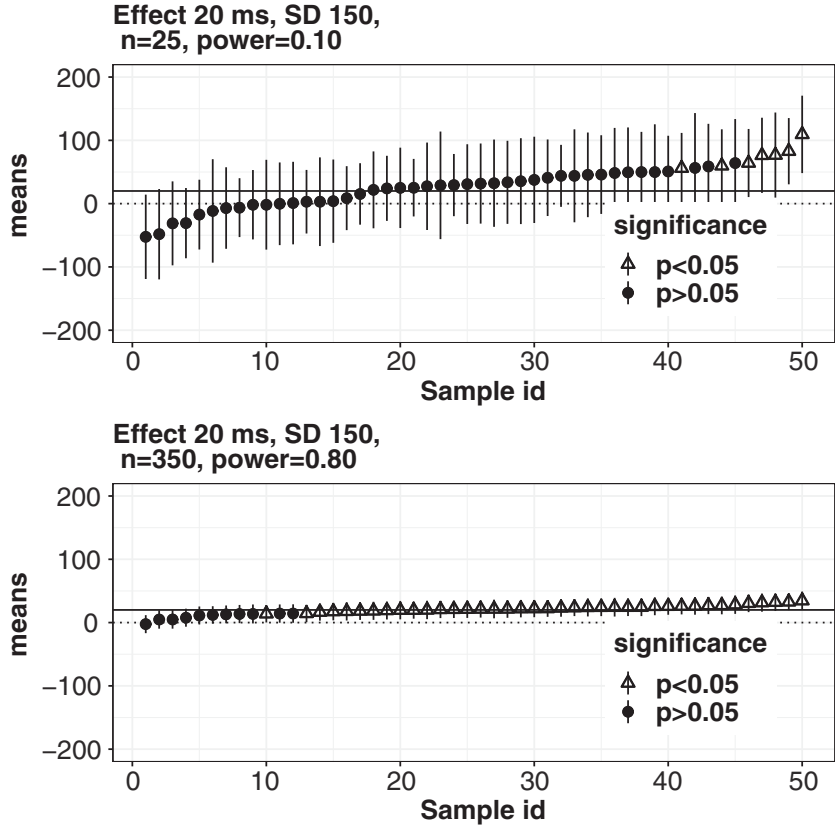

Figure 1.2 A demonstration of Type M and S error. Low power studies will yield overestimates and/or incorrect signs whenever a result is significant.

The Type M/S error issue is not just a theoretical statistical point; it has real practical consequences. For example, consider the eyetracking studies reported in Levy and Keller (2013). These studies claim to show evidence for surprisal effects (Hale, 2001; Levy, 2008), but seven replication attempts, including one higher powered study (100 participants vs. the original 28 participants) consistently failed to reproduce the claimed effect (Vasishth et al., 2018). It is quite possible that many such underpowered studies form the basis for theory development in psycholinguistics. We return to this point later when carrying out model evaluations on published interference effects.

Apart from the incorrect inferences that arise due to Type M/S error, another major problem in psycholinguistics is statistically incorrect inferences based on null results. Null results under repeated sampling can only be interpreted

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

if there is a demonstration of sufficient statistical power (Hoenig and Heisey, 2001) computed before conducting the studies. In the past, power has never been considered in such studies, but the situation has improved in recent years (e.g., Stack et al., 2018). This point about null results in low-power experiments is demonstrated in the upper part of Figure 1.2. If a researcher were to run an experiment with 10% power repeatedly, they would usually get a null result. Accepting the null result would be a mistake here, because the true estimate is not zero; it is just impossible to detect accurately using statistical significance as a criterion for discovery. Such incorrect inferences are quite common in psycholinguistics. The problems with such misinterpretations have been brought up repeatedly in the psychology literature (e.g., Cohen, 1962). But these problems with incorrect inferences from low power studies have generally have been ignored; a likely reason for this misuse of statistical theory is the cursory statistical education usually available in psycholinguistic curricula.

The Freedman-Spiegelhalter Approach In Bayesian approaches to clinical trials, an approach for evaluating predictions exists that is closely related to the Robert and Pashler criteria discussed above.[2] Simply put, the proposal is to posit a range of predicted values and then compare the estimates from the data with this predicted range. This method is discussed in Freedman et al. (1984) and Spiegelhalter et al. (1994). In recent years, this idea has been re-introduced into psychology by Kruschke (2014) as the region of practical equivalence (ROPE) approach.

The essential idea behind interpreting data using a ROPE is summarized in Figure 1.3. Assume that we have a model prediction spanning $[-36, -9]$ ms (Jäger et al., 2020). If we run our experiment until we have the same width as the predicted range (here, $36 - 9 = 27$ ms), then there are five possible uncertainty (confidence) intervals that can be observed; see Figure 1.3. The observed interval can be:

A. to the right of the predicted interval.
B. to the left of the predicted interval.
C. to the right of the predicted interval but overlapping with it.
D. to the left of the predicted interval but overlapping with it.
E. within the predicted range.

Only situation E shows complete consistency with the quantitative prediction. A and B are inconsistent with the model prediction; and C and D are inconclusive. There is a sixth possibility: one may not be able to collect data with the desired precision, and in that case, the observed interval could overlap

[2] The following section is from Vasishth and Gelman (2019), which is available under a CC-BY 4.0 Attribution International license.

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

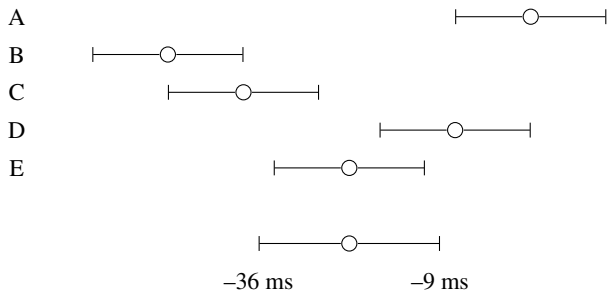

Figure 1.3 The five possible outcomes when using the null region or "region of practical equivalence" method for decision-making (Kruschke, 2015). Outcomes A and B are inconsistent with the quantitative predictions of the theory; C and D are inconclusive; and E is consistent with the quantitative theoretical prediction. Figure reproduced from Vasishth and Gelman (2019).

with the predicted range but may be much wider than it (here, the width of the predicted range is 27 ms). That would be an uninformative, low-precision study.

In contrast to the ROPE approach described above, what models in psycholinguistics usually predict is the sign of an effect, but not the magnitude or the uncertainty. This is one reason why null hypothesis significance testing is so popular: the question whether an effect is "present" vs. "absent" is easily answered by looking at the *p*-value.

But a prediction like "the effect is present" is not particularly useful because this implies that an effect with estimated mean 500 ms that is statistically significant would be consistent with the prediction just as well as a significant 5 ms effect. However, a 5 ms effect may have no special relevance for theory development.

If one really insists on using language like "the effect is present/absent," a more conservative way to proceed is to run the study until a Bayes factor of 10 (either in favour of the null or the alternative) is achieved. This is generally a more conservative approach than using the *p*-value with a cut-off of 0.05 because the bar for drawing a binary conclusion is much higher. Here, the fact that Bayes factor can be highly sensitive to the prior specification complicates the interpretation of a Bayes factor-based analysis. See Nicenboim and Vasishth (2016); Nicenboim et al. (2021) for discussion of the underlying theory of Bayes factors and its application in psycholinguistics.

1.6 The Goals of This Book

This book is a first attempt to address the gaps discussed in the previous section from a very particular perspective. In the following pages, we will spell out

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

a theory of sentence processing (Lewis and Vasishth, 2005) that uses (or is inspired by) a specific cognitive architecture, ACT-R (Anderson et al., 2004), which has been designed for modelling general cognitive processes. ACT-R is a reasonable choice for a framework because it is a mature architecture that has been widely used in artificial intelligence, human-computer interaction, psychology, and other areas of cognitive science to model human information processing (for examples, see the literature listed on the home page for the architecture: http://act-r.psy.cmu.edu/).

It is important to keep in mind that this is not a book about sentence processing models in general, but about one particular class of models, those relating to retrieval processes. The reader should therefore not be surprised to find that the discussion focusses on one kind of modelling approach.

1.6.1 Providing Open Source Model Code

One of the great failures of psycholinguistics has been that model code, as well as experimental data, are rarely made public (there have been, of course, some exceptions, particularly in recent years). A major goal of this book is to help change this unfortunate culture. This book can be seen as providing the reader with a complete report of all the modelling work on retrieval processes that we have done between 2005 and 2020. All the code and data associated with the modelling reported here can be reproduced by the reader, and extended and used to test novel predictions of the models. Of course, we recognize that code rot – the slow deterioration and increasing unusability of software over time – is inevitable. It is highly likely that five years from now the code provided will not work as expected, unless the reader can obtain the ACT-R, lisp, R, and Stan versions we used in creating the models. Even if the code fails to run some years from now, our hope is that at least the code can be adapted or rewritten to reproduce and extend the models presented here.

If the reader intends to run the code in this book, they should install ACT-R 6.0 and the R packages indicated in the source files for the book available from the GitHub repository (https://vasishth.github.io/RetrievalModels).

1.6.2 Modelling Average Effects as Well as Individual Differences

The book will discuss the modelling of both average effects and of individual differences. Specifically, we will illustrate how we could investigate (a) the influence of individual differences in working memory capacity on parsing; (b) the role of parsing strategy, including task-dependent underspecification; (c) the interaction between individual working memory capacity, grammatical knowledge, and parsing; (d) the interaction between the eye-movement control system and sentence comprehension; and (e) how individual-level differences

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

in the behaviour of individuals with aphasia might be explained in terms of model parameters.

1.6.3 Developing a Set of Modelling and Empirical Benchmarks for Future Model Comparison

A further goal of the book is to provide the next generation of researchers with a synthesis of the modelling and empirical work that followed the publication of the article by Lewis and Vasishth (2005). Our hope is that others will be able to build and improve on the present work, either falsifying or extending the empirical support for the model claims and thereby advancing our understanding of the important open theoretical issues, or developing competing models that can outperform the ones presented here. Some attempts at developing competing models that aim to outperform the models presented in the present book already exist (Cho et al., 2017; Parker, 2019; Rasmussen and Schuler, 2017; Smith et al., 2018). One problem common to all these models is that they take up one or two empirical phenomena of interest (a common choice is subject vs. object relatives, usually in English). This often leads to overfitting the model to a very narrow set of facts. A remarkable number of modelling studies (including our own!) limit themselves to narrow topics like relative clause processing. What is missing in the field is a set of benchmark empirical tests that a model can be evaluated on in order to demonstrate superior fit to data, relative to some baseline model. As a first step towards developing such a benchmark, we provide in one place the data-sets from reading studies on interference effect that happen to be publicly available.

Because the modelling reported here was carried out over many years (most of the work was done between 2005 and 2020), many computational challenges had to be overcome. For example, the ACT-R architecture itself is continuously evolving independently of the sentence processing architecture we work with. These version changes in ACT-R necessitated a near-complete rewrite of the modelling constructs. As a consequence, the original Lewis and Vasishth model's underlying machinery also changed in subtle ways. This evolution of ACT-R will remain a challenge for future researchers. A further problem that we encountered was that lisp is not a widely used programming language anymore; this makes the ACT-R model less accessible to researchers interested in using it. Fortunately, researchers have recently developed viable alternative implementations in python (Brasoveanu and Dotlačil, 2020), which may be easier to maintain and develop further. The lesson to be learnt here is that developing a sustainable code base, and preventing code rot, is a major challenge in any large programming project like this one, and the user/reader needs to be aware of this limitation and to be patient when adapting or using legacy code. One insight to be gained here is that perhaps some compromises

are necessary in order to make the theoretical machinery more accessible to the wider community. It is possible to investigate the core principles of the model without implementing a fully fledged model; this can be done by using code written in the programming language R. We provide such an implementation, along with a Shiny app that allows the reader to compute simple effects without doing any coding at all.[3]

Another issue we faced was that we studied different research problems piecemeal. For example, a model integrating eye-movement control and parsing is reported in Chapter 5, but this model has not been regression-tested with the core phenomena discussed in Chapter 4 or other chapters. Future generations working on this framework could (and should) develop a more systematic testing framework, so that empirical coverage is incremental in the sense that the model's performance on all previously modelled data-sets and phenomena is evaluated again when exploring an extension of the architecture.

One further area where the present work fell short was that modelling should ideally always be comparative; a baseline model is necessary to evaluate a particular model's relative performance. In more recent work, reported in Chapters 7 and 8, we have attempted to shift the focus towards evaluating competing models' performance on the same data-set. In future work, this should be standard practice. For example, an alternative competing model of eye-movement control and parsing would be very useful in order to better understand the relative performance of the model presented in Chapter 5.

Regarding the data used in this book, we focus almost exclusively on reading data, from self-paced reading or self-paced listening, and eyetracking studies. This is because we primarily set out to model the reading process; an exception is the modelling of visual world data reported in Patil et al. (2016a). We chose reading times as a convenient starting point because the Lewis and Vasishth (2005) model delivers predictions in terms of retrieval time and retrieval accuracy, and dependent measures in reading studies (e.g., fixation durations, comprehension accuracy) map relatively straightforwardly to this measure. We will therefore not discuss the large body of research using other methods such as electroencephalography (EEG) and the visual world paradigm. It is of course important to develop computational models that can be related to data that come from these methods. We hope that future generations will take up that task.

1.7 Looking Ahead

It may be useful to briefly summarize the structure of the remainder of this book. Chapter 2 reviews the range of empirical phenomena that form the basis for a large chunk of the modelling, and discusses the published empirical findings

[3] https://engelmann.shinyapps.io/inter-act/

https://doi.org/10.1017/9781316459560.006 Published online by Cambridge University Press

regarding these phenomena. One of the key takeaways from this chapter is that published studies on these phenomena are likely to be underpowered and therefore not sufficiently informative. Chapter 3 presents the core ACT-R model, as developed in the Lewis and Vasishth (2005) paper; Chapters 4 and 5 summarize two recent extensions. The first extension (Chapter 4) modifies the core model to account for linguistic prominence of items in memory and for so-called multi-associative cues. The second extension (Chapter 5) integrates an eye-movement control model (EMMA, a simplified version of the E-Z Reader model) with the parsing model and evaluates its performance. Chapter 6 then presents an evaluation of the model incorporating eye-movement control on psycholinguistic data on reanalysis and underspecification effects, and shows how individual differences in capacity can be explained by the model. One important question that needs to be answered is: how does the Lewis and Vasishth model fare in comparison to a competing model of retrieval processes? This is the topic of Chapter 7, which covers a model evaluation, using an implementation of the direct-access model of McElree as a baseline. Finally, Chapter 8 discusses the model's ability to explain individual-level differences in deficits in sentence comprehension in aphasia. The concluding chapter takes stock of the achievements and limitations of the work presented in the preceding chapters.