ILLC

A Study of Optimality Theory and the Human Sentence Processing Mechanism

By: Rajvinder Singh

Acknowledgements

The year 2001 was barely a month old when began the most trying period of my entire life. In August of that same year, I made the journey from Toronto to Amsterdam to embark upon a new (academic and otherwise) chapter in my life story. For making this journey as wonderful as it has been, many people are owed many words of thanks. I will use this section to 'name names'. I imagine that this section will be rather longer than most MoL 'Acknowledgement' sections. The reader who is easily put off by sentimental passages is advised to flip past this section to the 'meat and potatoes' of this document.

First, there was Toronto…

I must begin by pointing out that if it weren't for the unending love, sacrifice, and dedication that my mother Parminder Pal has shown me, I would not be here in Amsterdam. In fact, I don't know where I'd be, but certainly in much worse condition than I am in now. Mom, you have time and time again gone above and beyond the call of a mother's duty. I don't know how you did it, or where you kept getting the will to fight on, but you did. I don't know what better way to thank you than to ensure your efforts have not been in vain.

To my brother, Aminder Singh Multani, way back in 1985, little did I know that this kid who was stealing all my attention would grow up to be such a strong and mature young man. You have had to grow up faster than you should have, and you have dealt with all the challenges that have come your way with grace and dignity. You have already gained more wisdom than most people I know that are double your age. I guess it's easy though, with a brother like me! You have made me very proud.

To my 'other mom' Arvinder Kaur Padda, you have been what many would call my 'guardian angel' since as far back as I can remember. When it seemed like the whole world was walking out, you walked in, and refused to leave! I want to take this opportunity to thank you for all your help, your support, and your love. When I get back to T.O., we'll celebrate your '18th'' birthday by singing 'Boogie boogie boogie boogie dancing shoes…'.

To Harpreet Singh Padda, my cuz, second in command in IDA, all I've got to say is: I know that Apple's always got Walnut's back, no matter what. And vice-versa. Although you're younger than me, I look up to you in many ways, especially your open heart. It is quite well known that you and I have been tag teaming it since day one. That will never change. Ever. I am lucky to have you in my life.

To Rajdeep Kaur Padda, my other cuz, I hope that by the time this is printed, you have not been hooked up with some 'ref' on your recent trip to the I-Dot! But let me say right here, right now, that I am so proud of the person you have grown up to become. Your courage in very difficult times is something to be admired. Thanks for all your calls and emails, they always made my day (or night). But did you have to give Apple my contact info?

To my uncle Swinder Pal Singh Multani, I would like to express my gratitude for accompanying my mother and I on our trip to India in December (2001). Your support was much needed and much appreciated.

To my friend Sadat Anwar, you are amongst my oldest memories. Ever since we shared our first day of school *ever* in Ms. Rowsell's junior kindergarten class at Bala Avenue Junior Public School (cheap pop!), we have shared so many experiences that it would take another 22 years to recount them all. Some of the venues that felt the wrath of the Pretty Boyz include: Bala, CR, Weston, U of T, Stall #3, Stall #4, Gerrard, Tabaq, Albion, Woodside, Massey Hall, Maple Leaf Gardens, The Air Canada Centre, and the granddaddy of 'em all, (in my best Vinnie Mac voice), SKYDOME! Up next: m.W.o.

To my boy Harpreet Singh Dhillon, mad props for joining me on that crazy trip to Miami. I know that you did it just, if I may go street for a brief moment, 'to represent'. I will never forget it. I know that it must be difficult to be the 'Marty Jannetty' of 'the kliq', but you seem to be doing a fine job.

Last, but certainly not least, I want to thank my girlfriend Kamaljit Kaur Brar for…there is too much to state in this short space. This has been, in every respect, the most challenging test of our relationship, and at the end of the road, when all the dust has settled, we are still together, hand in hand. This is, in large part, due to the patience, understanding, and unlimited reservoir of love that you have showered me with. I know that it was very difficult at times. I want to say now something that I don't say enough to you: Thank you.

On the academic side of things, I was very lucky to have received an undergraduate education in a most stimulating environment at the University of Toronto. My professors there have given me VERY big shoes to fill. I will be very satisfied if I can make it up to the children's sizes. For starting me on the road to logic, language, and computation, I wish to thank the following wonderful teachers: Peter Apostoli, Jim Brown, Steve Cook, Ian Hacking, Alasdair Urquhart, and John Vervaeke.

…and then there was Amsterdam

Thanks must go, first and foremost, to my thesis supervisor Henk Zeevat, who provided me with expert supervision that I was all too pleased with. Henk, you read so many versions of my work, critiqued them all, praised them all, and ultimately always managed to get me to improve them. You also took the time to meet with me on many occasions. We seemed to have a penchant for disagreeing about something during every single meeting we had, but somehow, you always, always, sent me away knowing more than I knew before our meeting. That is what I call progress. I have developed as a scholar along many dimensions since my arrival here at the ILLC, and I owe a large part of that to you. Thank you for all your teaching, your letters, your discussions, your reviews, your ideas, and, most importantly, for your open heart. I always felt like you were looking out for my best interest. Thank you.

I would like to take this opportunity to thank Dick de Jongh for always having an open door to discuss any and all issues. I am especially appreciative of his appearance at my defence. Although he had prior (quite important) engagements, he made it out to see me 'go'. It meant a lot to me to have you there, Professor de Jongh, and I would like to express my gratitude to you.

To Ingrid van Loon, whose door was also never to be found shut on a student. You answered SO MANY of my questions before I got to Amsterdam (ranging from semi-intelligent to down right silly), and answered even more when I arrived here. Somehow, someway, you manage to keep this place together, running smooth like a US cruise missile, but for good, not destructive purposes. You deserve an award. I mean it.

I must thank the people responsible for allowing me to share a piece of the Spinoza Grant pie. If it weren't for the financial support I received, this majestic experience would not have been possible. I hope that my work has proven worthy of your judgement.

To many great teachers from whom I have learned so much, thank you for sharing your (immense) knowledge with me: Maria Aloni, Karen Kwast, Robert van Rooy, Domenico Zambella, and Henk Zeevat (again). Thanks also to Peter Paul de Witte for taking care of so many of my administrative problems.

To the members of my thesis committee, Remko Scha and Reinhard Blutner, your questions and comments were insightful, challenging, and very fair. They caused me to think about aspects of my work that had not occurred to me, and will serve to improve any extensions I make to the document that follows this sizeable acknowledgement section.

The ILLC would not have been such a rewarding experience were it not for the fact that I was blessed with having an absolutely terrific group of co-students. Thanks for all the good times to: David Wood, Steffen Bauer, Luciano Buratto, Seth Cable, Marian Counihan, Irwin Lim, Bernadette Hernandez, Will Rose, Joshua Sack, Börkur Sigurbjornsson, Wei Wei, Roger Antonsen, Mingzhong Tao, Ravi Rajani, Fabrice Nauze, Anna Pilatova, Balder ten Cate, Katrin Schulz and Marie Nilsenova. Special thanks to Seth, Ravi, and Luciano for many stimulating discussions on many stimulating topics. A very special thanks to Seth for reading (and critiquing!) my work, and allowing me the chance to do the same with yours. To Luciano: Keep fighting the good fight. It will, nay, it *must*, succeed.

To the New Kids on the Block, I enjoyed my time getting to know you (some better than others). Interesting times were had, and that logic party will go down in history as one of the craziest ever. To: Loredana Afanasiev, Guillame Aucher, Elizabeth Birchill, Jill Girasella, Spencer Gerhardt, Julia Grodel, Tanja Hotte, Gilad Mishne, Clive Nettey, Thuy Link Nguyen, Oren Tsur, Chunlai Zhou, Dirk Walther and Andreas Zollman, if I may pass on a piece of advice: One of the best resources at your disposal is each other. Do not hesitate to use it.

Away from the world of the ILLC, I had the pleasure of meeting a great number of people from all over the globe. To all my peeps (and freaks) from Valkenierstraat: You guys brought a great deal of joy into my life (as much as I can remember, anyways…). Je me souviens.

Finally, I had the distinct pleasure of living in the middle of nowhere with an amazing group of guys. We bonded very quickly, very well, and Kruislaan soon became one big happy, crazy, argumentative family. I am certain that the period beginning 1 September, 2001, and ending 31 August, 2002, was the greatest in Kruislaan 306/308 history, and will never again be repeated. Highest thanks go to: Ferdinand 'Bottle' Alimadhi, Kostas 'Nobel' Sakkos, Adianto 'Cafu' Wibisono, Sushil 'Ronaldinho' Adhikari, Ilia 'Turing/Minsky' Korjoukov, Qiu 'Figo', and Changhong. A special thanks to Ilia for his much needed help in October. I would also like to thank Simolappi and CJ Faux, adjunct Kruislaan members, for some very good times. Further thanks to Simon 'Pieman' Engler for opening the doors of Djambistraat (and hence the wonderful world of junkies, one of whom Ilia and I had the distinct pleasure of 'catching'…) and being a really cool bud all the way through.

Gone but not Forgotten

In this past short time, I have lost many close people to death. All were very young, and all have left behind a great number of people that will forever be saddened by their loss. To my uncles Jagjit Singh Multani, Santokh Singh Padda, and Manjit Singh, I would like to say that your families are very strong, they are not alone, and that we, together, will get through any and all problems. Our memories of you bring us joy, the bond we shared with you is stronger than ever, and the strength and courage that you always showed in your lives lives on in our veins, pumping us through each and every day. You all impacted my life in very powerful ways, and I carry a piece of you each and every second of each and every day.

We at Kruislaan also lost our very good friend Fan Chen to a very surprise attack of cancer. Fan Chen was always a terrific friend, always ready to give sound advice, and was about as decent a human being as you can find. You are sadly missed, my friend.

On 26 April, 2001, I lost my father to a very long battle with a very serious disease. This was the most tragic moment in my entire life. I cry, Dad, perhaps not so much for your passing, for such is inevitable, but for your time spent here. Your name, so fitting, Kashmir Singh. Kashmir. India's paradise and India's downfall. You had all the talents in the world, intelligence, charm, the ability to adapt to new cultures, new languages, an ease with the ladies (that I picked up!), and yet, there was always trouble, there was always conflict, there was always something. That something took you in the end, but not before you passed on to me your many gifts. The one that was most precious to me, and still is, is your friendship. We were more like friends than we were like father and son. Perhaps the ultimate friendship is a father-son bond. I don't know. But I do know that I miss you dearly, that I wish I could have one more chance to hang out with you, and that you would be there at Toronto's Pearson International Airport to welcome me back home from completing a Master of Logic degree from the ILLC.

0. Introduction

From a computational perspective, parsing is a very interesting phenomenon. All people do it quickly, and all people do it well. The history of cognitive science has been filled with attempts to explain the mechanisms that guide the human sentence processing mechanism (hf. HSPM). These have typically made underlying assumptions about other cognitive faculties, such as working memory and modularity. Our work will be no different.

In what follows, we develop a theory of parsing that is grounded in the framework of optimality theory (hf. OT). The goals of this work may be explicitly stated as follows. First, we hope to construct a theory of the HSPM that is both descriptive and explanatory. This will require us to use both rational analysis and empirical data in order to enumerate all and only those constraints that are involved in the functioning of the HSPM. Second, we want to implement these constraints into an OT system. Before doing so, however, it is important to ask: Why should we want to marry parsing theory with OT?

There are two main reasons. First, the notion of 'optimality' has been assumed throughout the history of research on parsing. When faced with local ambiguities, listeners/readers must choose between a set of candidate structures, guided by interacting constraints. At an abstract level, this process of resolving syntactic ambiguities is almost identical to the process of determining grammaticality in optimality theoretic systems. At the very least, the structural similarity suggests that OT might prove to be useful in modelling observed parsing behaviour. Second, although OT's origins (and greatest successes) lie in the domain of phonology, it has recently undergone extensive expansion into the domains of syntax, semantics, and pragmatics. This expansion, coupled with its infamous narrowing of the competence-performance distinction, leads very naturally to the question, can OT encompass the domain of language processing as well? Part of our quest is to answer this question in the affirmative, thereby continuing the expansion of OT qua linguistic theory qua theory of the language faculty. However, we are not the first to attempt to merge OT and the HSPM.

Gibson and Broihier (1998, hf. GB) provide OT implementations of various prominent theories of parsing from the current literature. They translate three different constraint sets into OT systems. The first is the famous 'garden path theory of sentence processing', which consists of the constraints Minimal Attachment and Late Closure. The second contains constraints involving thematic role assignments and preferences to attach locally. The third set consists of constraints that indicate a preference for attachments that are local and near a predicate.[1] Unfortunately, it is found that none of the OT implementations are able to account for the data. GB use this to argue that 'standard OT', which is used here to mean OT with the property of strict domination[2], is unable to accommodate observed parsing preferences. They claim that a weighted constraint theory that allows lower ranked constraints to additively outweigh higher ranked constraints would yield greater empirical coverage. In this thesis, we take issue with GB's conclusion by attempting to demonstrate the effectiveness of 'standard OT' in accounting for the experimental data. To accomplish our task, it will be necessary to turn to the psycholinguistic literature to help guide the development of our system. We claim that the theories of the HSPM posed in GB are unsuccessful precisely *because* they fail to incorporate well known psycholinguistic results into their theoretical formalism, *not* because standard OT is not up to the task. By incorporating the experimental results into our OT system, we hope to have a more psychologically plausible theoretical construction that improves current descriptions of the HSPM. Upon developing our system and analysing its adequacy in describing and explaining the data, we will come back to the issue of whether or not standard OT is able to account for the observed parsing preferences.

[1] The first constraint set is outlined in: Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies.* Ph.D. Dissertation, University of Connecticut. The second is outlined in: Pritchett, B. (1988). Garden path phenomena and the grammatical basis of language processing. *Language* 64: 539-576. The third is outlined in Gibson, E., N. Pearlmutter, E. Canseco-Gonzales and G. Hickok. (1996). Recency preference in the human sentence processing mechanism. *Cognition* 59, 23-59.

[2] The property of strict domination says that, for any two constraints C and D such that C is ranked higher than D (notationally represented as C >> D), no number of violations of constraint D is as destructive as a single violation of constraint C. For example, if two candidate representations X and Y are such that X incurs no violations of C and five violations of D, whereas Y has one violation of C and none of D, and neither X nor Y violate any other constraints, then X is more optimal than Y. This is also represented as X >> Y.

So, here is the overall structure of our story. We begin in Chapter 1 with a gloss of the most influential account of parsing to date, viz., the 'garden path theory of sentence processing'. Almost all of the research in language processing has been a response to the garden path theory (hf. GPT). Each response has served to either contribute to the GPT or to provide a critique of its axioms and theorems. Indeed, the psycholinguistic literature we examine is a debate on the status of the constraints making up the GPT. Thus, in order to appreciate the experimental results examined in this work, it will be necessary to have the GPT as the backdrop against which we interpret the psycholinguistic data. Hence, we begin with a brief overview of the GPT.

In Chapter 2 we examine a broad range of psycholinguistic results with two goals in mind. First, we will use the experimental data to test the adequacy of the GPT. Second, by analysing the GPT's successes and, more importantly, its failures, it is hoped that important insights will be revealed that will guide us to a more accurate theory of the HSPM.

The experimental results reveal that, indeed, the GPT is flawed in fundamental respects. First, it includes constraints that are not 'doing anything' in that they do not seem to be involved in the determination of parsing preferences. Second, it is missing certain constraints that are necessary to capture the data. These results can be thought of as remarks on the 'soundness' and 'completeness' of the garden path theory.[3] Suppose G is a theory (constraint set) for a particular domain X (eg. Language processing). Suppose further that some subset of G may be expanded (by adding constraints) to the 'right' theory T of domain X.[4] We say that G is 'sound' if $G \subseteq T$. We say that G is 'complete' if

[3] According to standard usage, grammars are theories, and hence are subject to adequacy conditions. Here, we propose 'soundness' and 'completeness' as two such conditions. I use scare quotes only because these terms are also used in the logical literature as remarks on logical theories. Although there are some clear parallels between the way the terms are used in the well-established logical tradition and the way they are used here, I do not want to confound the two. Hence, as a sign of respect to the logical tradition, I use 'scare quotes' to highlight the fact that these terms are not new, are not unique, and are borrowed from a rich tradition, and would like to note that the usage has been modified.

[4] Of course, this is abstracting away from the fact that for any given data set, there is an unbounded number of theories that can correctly describe that data set. The discussion here assumes that we are 'within' a Kuhnian paradigm, and that, within this paradigm, there is some such 'right' theory T that G may or may not be close to. The 'soundness' and 'completeness' of G are measures of how well it approximates T. This will not please philosophers, but we do not import any realist assumptions here. We are simply

G $\supseteq$T. According to these definitions, G will be sound only if it does not contain any superfluous or vacuous constraints. G will be 'complete' only if it includes all the constraints that are involved in some domain X. In Chapter 2, we will find that the GPT is not 'sound' because it includes constraints that are not reflective of the actual constraints guiding the HSPM. It is not 'complete' because it is missing constraints that are necessary to explain the computations of the HSPM.

In Chapter 3, using these results as our guide, we implement the constraints that *are* involved in parsing preferences into a very natural OT system. The constructed system is a standard OT system whose constraints satisfy the property of strict domination. This is important, for, contra GB we are attempting to demonstrate that standard OT is well-suited to the task of predicting parsing data. The system we develop has many nice properties. First, it consists of a small set of constraints that carry both rational and empirical support. Second, the constraints are clearly motivated by the need to make the HSPM computationally efficient, which we consider to be its most pervasive feature. Third, not only does the theory incorporate well-known psycholinguistic results into its formalism, it makes very explicit the nature of the cognitive architecture that allows the observed psycholinguistic phenomenon to take place at all. This takes our work farther than most other theories of the HSPM, for they often have little to say about the architectures that allow the cognitive phenomenon they describe to arise in the first place (but see Lewis 2000). By being precise about both the architectural and computational properties of the phenomenon under consideration, we are able to make clear, falsifiable predictions. Our theory is further constrained by the fact that architectural assumptions and computational assumptions mutually restrict each other; certain kinds of architectures rule out (and imply) certain kinds of computations, and certain kinds of computations rule out (and imply) certain kinds of architectures.

In Chapter 4, we test the adequacy of the system by comparing its predictions with the observed data. We demonstrate that it is able to capture a large array of experimental

---

making terminological definitions so that we may discuss the 'goodness' of theories in the processing literature, such as the GPT.

results, predicting observed parsing preferences in English and Spanish ambiguities. Additionally, the system is able to predict differences in the processing complexity of unambiguous sentences where there are no 'preferences'. For example, it is well known that centre-embedded structures are more difficult to process than right (and left) branching structures. Our system is able to predict this difference in processing complexity.

One of the factors contributing to the theory's descriptive power is that the notion of structural ambiguity becomes much streamlined in our work. We illustrate that a set of structural ambiguities previously thought to be unrelated can in fact be reduced to a smaller set of ambiguity types. This result follows almost directly from our architectural assumptions. We demonstrate that there is actually a redundancy in the kinds of ambiguities faced by the HSPM, allowing it to repeatedly use general resolution methods rather than construction specific mechanisms to resolve the ambiguities it comes across. This serves to add efficiency to the HSPM's computations and adds a touch of elegance to the theory being developed.

In Chapter 5, we offer some remarks on the impact of our work on broader issues in linguistic theory, theories of the HSPM, and cognitive science in general. These include, inter alia, discussions of topics such as the relation between the parser and the grammar, general cognitive architecture, language acquisition, and the adequacy of standard OT as a framework for language processing. Such discussions will also be found interspersed throughout the text, often made to help motivate or justify various assumptions or conclusions that we make or draw. Ultimately, we hope to have developed a sophisticated theory of the human sentence processing mechanism that is descriptively powerful, theoretically sound, and consistent with what we know about human psychology. Furthermore, by providing a successful translation of the theory of the HSPM into an OT constraint system, we hope to expand OT to encompass linguistic performance in addition to its coverage of linguistic competence. As mentioned at the outset, the determination of parsing preferences and the determination of optimality in OT are structurally remarkably similar. Hence, the prospects for a successful merger

between OT and language processing are a priori quite promising. This work examines the extent to which these prospects may be formally achieved.

Enough with the introductory remarks! A story is waiting to be told, and so to the garden paths we go.

1. The Garden Path Theory of Sentence Processing

The garden path theory of sentence processing (hf. GPT) is a collection of ideas attempting to describe the constraints that guide sentence comprehension. For many sentences there are several possible interpretations of their syntactic and semantic structures. However, in the normal course of events, only one is selected, often out of a very small competing set. How is our interpretation so constrained? The idea behind the GPT is that the HSPM makes life simple for itself by minimising processing costs at each stage of the parse. This is formally cashed out via the 'immediacy principle', which underlies the two constraints making up the GPT, viz., the 'Late Closure Strategy' and the 'Minimal Attachment Principle'.

1.1 Immediacy Principle and Garden Paths

Carroll (1999) argues that in processing, we use a heuristic called the 'immediacy principle'. As the name suggests, the immediacy principle states that the decision about where to place incoming words into the phrase marker being constructed is made immediately upon encountering a word. The HSPM is building the syntactic structure 'on-line', and must make such decisions as soon as possible. The advantage of using this technique is rather obvious. If the HSPM were to wait to see where the sentence was heading before interpreting a particular word or phrase, the amount of information to be activated and held would quickly overwhelm working memory. For consider a sentence with three choice points, or points of ambiguity. Suppose further that at each such point there are three possible options for interpretation. The hearer (or reader) would thus need to hold nine different sentences in her head in order to interpret such a sentence. This seems highly unlikely, given the well-known limitations of working memory. It is important to note right from the outset the importance of considerations external to the

language faculty in the development of constraints within the language faculty. In this case, working memory limitations drive the need for the HSPM to interpret words immediately, rather than to take a 'wait-and-see' approach.

Of course, as with every advantage, there is sure to be a disadvantage. In this case, the disadvantage is that by virtue of using the immediacy principle the HSPM often makes mistakes in parsing decisions. After proceeding along a particular path in the parse, the HSPM realises it has incorrectly interpreted a particular word or phrase, and must therefore reanalyse the input in order to arrive at the correct interpretation. Such misparsed sentences are called 'garden path sentences'. Due to an error having been made in their initial interpretation, a clear prediction can be made: Garden path sentences should be difficult to process. This processing difficulty is reflected through increased time complexity in comprehension tasks. There is a whole cottage industry's worth of examples of such sentences. Popular ones include[5]:

(1) The horse raced past the barn fell.
(2) The florist sent the flowers was pleased.
(3) Since he jogs a mile seems like a short distance to him.

1.2 Late Closure Strategy

The Late Closure Strategy states that, whenever possible, incoming lexical items are attached to the current constituent (Carroll, 1999). Note that, as per the immediacy principle, the Late Closure Strategy works to reduce the load on working memory. The

[5] I have a problem with sentence (3) being considered a garden path sentence in the literature. The reader is not led up the garden path because of structural ambiguity, but rather because the writer has deliberately made it difficult for the reader to interpret the sentence. There should be a 'comma' after the word 'jogs'. In spoken language, for example, the sentence would be disambiguated by suprasegmental cues, such as intonation, pause, rate of speech, etc. My thesis advisor Henk Zeevat and I both agree that there is an important connection between prosody and comprehension. However, I have, as yet, no substantial account of what that connection is, and would like to reserve that line of reasoning for future work. As a result, the account developed here will be incomplete. The long and short of this long footnote is that sentences such as (3) are not really garden path sentences, for any such sentence must assume that writers/speakers are following Gricean maxims of communication, faithfully representing their thoughts in as clear a manner as possible. Sentence (3) does not satisfy such a precondition, and as such should not be considered a true garden path sentence.

force behind the strategy is best explained with an example. Consider the following sentence:

(4) Tom said that Bill had taken the cleaning out yesterday.

Here, 'yesterday' can be interpreted as modifying the main clause 'Tom said…' or as modifying the subordinate clause 'Bill had...'. Late Closure resolves the ambiguity by positing a preference for the latter alternative.

Note that the theory is well constrained in the sense that it makes very clear, falsifiable predictions. One prediction is that when faced with ambiguities of the 'local attachment' versus 'non-local attachment' type, the HSPM will prefer to attach locally. Another prediction follows from this: Because parsers prefer to attach locally, sentences that force non-local attachment should take longer to process than sentences that force local attachment (where such attachment preferences can be forced by disambiguating the sentence via number, gender, tense, etc.).

1.3 Minimal Attachment Principle

The Minimal Attachment Principle states that parsers prefer to attach incoming words into the phrase marker with the minimum number of nodes consistent with the well-formedness rules of the language (Carroll, 1999). Note the relationship between Minimal Attachment and working memory: more nodes translate into more processing cost. Thus, the HSPM prefers interpretations requiring as few nodes as possible such that the resulting structure is still well-formed. To illustrate, consider the following sentences:

(5) Alfons ate the soup with tomatoes.
(6) Alfons ate the soup with a spoon.

Initially, one attempts to attach the PP to the VP rather than to open a new node for NP attachment.[6] In (6) this works fine. However, in (5), the thematic processor does not accept the syntactic output. Hence, a re-alignment is necessary, and a new node must be opened.

The theory is once again shown to make clear, precise, falsifiable predictions. For example, in PP attachment ambiguities like (5) and (6), NP-attachments (as in (5)) should take longer to process than VP-attachments (as in (6)).

1.4 Further Aspects of the System

The above presentation suggests that the HSPM is modular, with syntactic processing taking priority over semantic and pragmatic processing. For example we argued that in (5) the thematic processor receives (as input) a syntactic output, and subsequently rejects that (syntactic) output. This suggests that the syntactic processor works independently without interference from other linguistic/cognitive domains. It functions in a way that is entirely informationally encapsulated. There is much evidence against this view, however, which we shall present in Chapter 2. To foreshadow somewhat, some candidates for interference with syntactic constraints are frequencies, context, and real-world knowledge. We shall have the opportunity to examine each of these candidates. Some emerge as absolutely essential in determining parsing preferences, while others are seen to have limited importance. The evaluation of which candidates play active roles in the parsing process will have a great impact on the development of our OT system.

In the parsing literature, it is generally held that in cases of conflict between Minimal Attachment and Late Closure, Minimal Attachment wins. For example, in sentences (5) and (6), at the word 'with', Late Closure prefers NP-attachment, whereas Minimal Attachment favours VP-attachment (cf. Footnote 6). In this situation of conflict, Minimal Attachment is posited to come out victorious, guiding the interpretation to VP-attachment.

[6] Of course, this depends on the kind of syntactic framework being used. We discuss the framework we are using a bit later in the text. See (11) and (12) in Section 2.1.1 for a similar sentence for which we provide a

Thus, we may characterise the GPT with the following basic tenets. First, the constraint set is {Minimal Attachment, Late Closure}. Second, Minimal Attachment dominates Late Closure. Finally, the parser is modular in architecture, where Minimal Attachment and Late Closure always apply to candidate structures and always apply first. In the next chapter, we make recourse to experimental results in order to test the adequacy of the GPT in accounting for parsing effects. By doing so, we will have a clearer idea of the relative strengths and weaknesses of the GPT. Using this information will give us insight into the actual mechanisms that the HSPM does and does not use. By examining the successes of the GPT, we simply incorporate its validated constraints into our model. Examining its failures will reveal which constraints should be avoided in our formalism, as well as give us insight into the directions our model should take.

2. Psycholinguistic Results

In order to arrive at an adequate theory of the mechanisms employed by the HSPM, it is necessary to combine empirical data with rational analysis. Experimental results allow us to test proposed theories by isolating hypothesised constraints and analysing their actual role in language processing. By shedding light on the underlying mechanisms that are employed in parsing, empirical results in effect help guide and constrain the development of our theories. In this chapter, we use experimental results to test the validity of the basic tenets of the GPT outlined in Chapter 1. The outcome of these investigations will be used to help us construct our theory of the HSPM in Chapter 3.

Recall the basic tenets of the GPT: the constraint set is {Minimal Attachment, Late Closure}, Minimal Attachment wins in cases of conflict with Late Closure, and the HSPM is a modular processor. This theory was dominant until the late 1980's, when a wealth of empirical results began to cast a large shadow of doubt on each of its basic tenets. In what follows, we will illustrate how each of these tenets is flawed. In Section 2.1, we investigate the axiom that Minimal Attachment dominates Late Closure, and conclude that it is unsubstantial. In Section 2.2, we outline important psycholinguistic

---

corresponding analysis, to clarify where the extra node for the NP-attachment comes from.

work illustrating that the constraint set C={Minimal Attachment, Late Closure} is neither 'sound' nor 'complete', as well as indicating that the HSPM is not informationally encapsulated. Section 2.2.1 is devoted largely to the important work of Taraban and McClelland (1988, hf. TM), which simultaneously reveals three pieces of information for us. First, it teaches us that the constraint set C={Minimal Attachment, Late Closure} is not 'complete', for it excludes a most important factor in the parsing process, viz., frequency information stored with lexical items. Second, it teaches us that the constraint set C is not 'sound', for it contains a vacuous subset of constraints. Third, it teaches us that the HSPM is not informationally encapsulated, due to the fact that it accesses frequency information from the lexicon. These results raise the following question: If the set {Minimal Attachment, Late Closure} is not 'sound', then which of the constraints is to blame for this property? We take TM's results to be on attack on both of the constraints. However, in Section 2.2.1, we illustrate that Late Closure has some independent evidence supporting its role in the HSPM. In contrast, Minimal Attachment does not have any such support. Hence, we will exclude Minimal Attachment from our constraint set, but (a variant of) Late Closure will remain. Finally, having seen that the HSPM is non-modular, we examine two factors that have been claimed to interact with syntactic constraints in the process of sentence comprehension, namely context and real-world knowledge. We will find that neither factor is involved in *initial* parsing decisions.

2.1 Minimal Attachment versus Late Closure

It is useful at this juncture to translate the constraint set {Minimal Attachment, Late Closure} into an optimality theoretic constraint system that provides absolute measures of the well-formedness of candidate structures. As currently stated, the constraints measure the well-formedness of candidate structures with respect to the well-formedness of others, thus making quantitative analysis difficult to come by.

We follow the OT system developed in GB. They translate Minimal Attachment and Late Closure into Node Conservativity and Node Locality, respectively:

(7) Node Conservativity: Don't create a phrase structure node.

(8) Node Locality: Attach inside the most local maximal projection.

This reformulation of the constraint system allows us to make absolute measures of how well candidate structures satisfy the constraints. We measure violations of Node Conservativity incurred by a candidate structure at a given point in the parse step by counting the number of phrase structure nodes that need to be created at that step. Violations of Node Locality are counted as the number of maximal projections on the right perimeter of the current structure that are passed over in making an attachment. GB translate this requirement more formally as follows:

> *An attachment to structure XP at the node Y in XP is associated with one locality violation for each maximal projection on the right perimeter of XP that is dominated by Y.* (GB, p.161)

Further assumptions built into the system are as follows.[7] First, the inputs are sequences of lexical items, such as (the), (the, horse), (the, horse, raced), etc. Outputs are well-formed phrase structures that extend outputs of earlier inputs. Let us call this property 'extensionality'. To illustrate the idea behind 'extensionality', consider the sentence 'The horse raced past the barn fell'. The lexical item (the) gets mapped to an output, say O1. Then, (the, horse) gets mapped to another output O2, where O2=(O1+x), and x extends O1. Note that O2 cannot reanalyse O1; at each stage in the parse, only those candidates are allowed that extend O1 without changing the previous 'path' of the parse.[8] Finally, words in the input cannot be skipped[9], evaluated candidates at any point in the parse are grammatical, there are no vacuous projections in the phrase structure grammar, and ternary branching is permitted.

[7] See GB, and Blutner (2001).
[8] This assumption about our PSG puts important constraints on the notion of reanalysis. Reanalysis will occur in our system by backtracking to the point of ambiguity and following an alternative path. In particular, the HSPM will not look for a more optimal candidate from the current output offered by GEN. It is assumed that the mistake in parsing occurred earlier in the parse; as the current output of GEN contains only extensions of previous parses, all the current candidates are also misparses. Hence, backtracking is required to repair the garden path effect.
[9] This constraint is inviolable, for otherwise a null structure with no nodes or branches would be optimal.

2.1.1 In Support of the GPT

One of the factors contributing to the longevity enjoyed by the GPT is that it accounts for a broad range of data. In the OT translation of the GPT, the ranking is as follows: Node Conservativity >> Node Locality. This ranking is able to describe a large set of parsing preferences. We illustrate with two examples.

Consider again sentence (1), 'The horse raced past the barn fell'. After processing 'the horse', the constructed structure of the sentence is: [IP[NP[D the][N horse]]]. The ambiguity arises, of course, at the word 'raced'. Here, it could be analysed as either the main verb of the matrix clause or as part of a reduced relative clause. The main clause analysis, along with its Conservativity and Locality violations, is shown in (9). The reduced relative analysis is illustrated in (10).

(9) [IP[NP[D the][N horse]][VP raced]]
Conservativity Violations: 1 (the new VP node)
Locality Violations: 1 (NP skipped and attachment made to IP)

(10) [IP[NP[D the][N'[N horse][CP[IP[VP raced]]]]]]
Conservativity Violations: 4 (the new nodes N', CP, IP, VP)
Locality Violations: 0 (because attachment is made inside NP, which is the most local maximal projection)

As Node Conservativity >> Node Locality, the system predicts that the parser will prefer sentence (9). Due to this preference, it is predicted that the HSPM will be forced to reanalyse the sentence upon reaching the word 'fell'. This is indeed found to be the case. Thus, the constraint ranking predicts the garden path effect observed with this sentence. Note that the ranking is necessary in this system of constraints in order to predict (9) >> (10), for (10) fares better than (9) with respect to Locality.

As a second example, consider ambiguities involving PP attachments, as we saw earlier in (5) and (6). We use a different set of sentences this time, modified from Taraban and McClelland (1988):

(11) Jap saw the cop with binoculars. (VP-attachment)
(12) Jap saw the cop with a revolver. (NP-attachment)

The crucial ambiguity arises at the word 'with', for the PP has two possible attachment locations. Until this point, the structure of the sentence is:

(13) [IP[NP Jap][VP[V saw][NP[D the][N cop]]]]

The VP-attachment is analysed in (14), and the NP-attachment is analysed in (15).

(14) [IP[NP Jap][VP[V saw][NP[D the][N cop]][PP with]]]
Conservativity Violations: 1 (new PP node)
Locality Violations: 1 (PP attachment to VP rather than more local NP)

(15) [IP[NP Jap][VP[V saw][NP[D the][N'[N cop][PP with]]]]]
Conservativity Violations: 2 (new nodes N', PP)
Locality Violations: 0

Psycholinguistic data reveal that sentence (11)) is processed faster than sentence (12) (Taraban and McClelland, 1988). The proposed ranking predicts this result, for the VP-attachment (14) incurs less violations of the higher ranked Conservativity constraint than the NP-attachment (15). Again, it is crucial here that Conservativity be ranked higher than Locality, for (15) is more optimal with respect to Locality.

Examples such as these provide support for the view that Conservativity >> Locality. Crucially, each of the above examples requires Conservativity to outrank Locality in order to predict the observed parsing preferences, for the preferred structure suffers more

violations of Locality than its competitor. However, there are numerous cases where any ranking between Conservativity and Locality will be able to account for the data. On its own, this fact does not pose a problem for the current hypothesised ranking for it still captures the observed parsing preferences. However, with the addition of data that require Locality to outrank Conservativity, the view championed by the GPT no longer remains tenable. There does not seem to by any systematic ranking, even within single languages, of the constraints Conservativity and Locality. Indeed, if no ranking is possible between them, as required by a standard OT system, then either: (a) Standard OT is insufficient as a theoretical framework; (b) (Elements of) The garden path theory are invalid; or (c) Both standard OT and the GPT are flawed in fundamental ways. In the next section, we provide a single example of a sentence where any ranking between Conservativity and Locality is able to account for the data. We follow this up in the subsequent section with an example that requires the ranking Locality >> Conservativity in order to account for the parsing patterns. In each of these cases, there are several more examples that illustrate the point. However, we demonstrate one example from each solely for illustrative purposes.[10] This will lead into further psycholinguistic results that will allow us to continue our examination of the tenets of the GPT. We will be assuming throughout that standard OT is well able to handle parsing, and that if there are any problems with empirical coverage, they derive from the GPT. In other words, we are assuming that (a) and (c) are both false, while (b) is true. Hence, in Section 2.2, we will continue the examination of the GPT, using psycholinguistic data to test its 'soundness' and 'completeness', as well as its assumption of modularity. But first, we must finish our investigation of the proper ranking between the two constraints Minimal Attachment and Late Closure.

2.1.2 The Ranking is Irrelevant

In this section we give an example of a sentence that does not require any particular ranking to hold between Conservativity and Locality; any permutation of the dominance relation will do. The case we consider is one where the verb subcategorizes for both NP and CP complements. The verb 'know', for example, can take both a simple NP, as in

[10] The interested reader may refer to GB for a wealth of such examples.

'Ashesh knew Randeep', as well as a sentence complement, as in 'Ashesh knew Randeep liked Samira'. It has been argued (Frazier and Rayner, 1982) with some empirical support that CP continuations are more complex than simple NP ones.[11] Thus, under this assumption, a prediction of any theory of parsing should be that CP continuations take longer to process than NP continuations. An OT system with the ranking Conservativity >> Locality is well able to capture this idea. After having processed 'Ashesh knew', the structure is as follows:

(16) [IP[NP Ashesh][VP knew]]

The ambiguity arises at 'Randeep'. Should 'Randeep' be interpreted as the direct object of 'knew' or as the subject of a CP argument of 'knew'? The structure in (17) shows the former interpretation, and that in (18) shows the latter:

(17) [IP[NP Ashesh][VP[V knew][NP Randeep]]]
Conservativity Violations: 2 (new nodes V, NP)
Locality Violations: 0

(18) [IP[NP Ashesh][VP[V knew][CP[C e][IP[NP Randeep]]]]]
Conservativity Violations: 5 (new nodes V, CP, C, IP, NP)
Locality Violations: 0

The system correctly predicts that (17) >> (18). However, note that this result is independent of the ranking of the constraints. The same result would hold if the ranking were reversed to Node Locality >> Node Conservativity.

2.1.3 Is a Ranking Even Possible?

[11] This generalisation is actually not valid, as we will see later in the paper. However, the present discussion is not disturbed by this fact, as the example being used here satisfies the prediction. The point of the present discussion is to examine the robustness of the ranking Conservativity >> Locality, which is argued here to not be as strong and robust as the GPT would have us think.

In this section, we provide evidence that the ranking Conservativity >> Locality is unable to account for parsing preferences. It will become evident that, on pain of losing much descriptive coverage, no ranking between the two constraints is possible. Upon presenting the support for this claim, we will delve deeper into the examination of the tenets of the GPT. We turn now to the examples.

Recall (cf. Section 1.3) the sentences (5) and (6). They involve an ambiguity concerning PP attachment, where both a VP-attachment and an NP-attachment are possible. Recall further that Minimal Attachment prefers VP-attachments over NP-attachments, due to the reduced number of nodes that are required to be opened. Thus, the GPT predicts that VP-attachments should be processed quicker than NP-attachments. Consider now the following sentences, taken from Taraban and McClleland (1988, hf. TM):

(19) I read the article in the magazine. (NP-attachment)
(20) I read the article in the bathtub. (VP-attachment)

The crucial ambiguity in each sentence occurs at the word ‘in’, where it must be decided where to attach the PP. The current OT system predicts that (20) is more optimal than (19). For observe that at the point of ambiguity, the structure of the sentence is:

(21) [IP[NP I][VP[V read][NP[D the][N article]]]]

If the parse proceeds as an NP-attachment at the point of ambiguity, the resulting structure, along with its constraint violations, is as in (22). The corresponding analysis of the VP-attachment is given in (23).

(22) [IP[NP I][VP[V read][NP[D the][N’[N article][PP in]]]]]
Conservativity Violations: 2 (new nodes N’, PP)
Locality Violations: 0

(23) [IP[NP I][VP[V read][NP[D the][N article]][PP in]]]

Conservativity Violations: 1 (new node PP)
Locality Violations: 1 (NP skipped in favour of VP as point of PP attachment)

Thus, at this stage in the parse, (23) >> (22) according to the GPT, and hence (20) is predicted to be processed faster than (19). However, TM find that the opposite result is the case: (19) is processed faster than (20). *Under the assumption* that Minimal Attachment and Late Closure are both at work, and are the sole constraints at work, the only way to predict this result is to have Node Locality outrank Node Conservativity.

2.1.4 Brief Discussion

The above results indicate that no ranking is possible between Node Conservativity and Node Locality. What are we to make of this? GB conclude that this result can be used to refute the use of strict domination in an OT theory of parsing. However, we argue that such a conclusion is not immediately warranted. It may instead be the case that the GPT is flawed, with the adequacy of standard OT left undamaged (cf. 2.1.1). Thus, we maintain here the running assumption that standard OT is well suited to account for parsing, and continue our examination of the GPT. We have already seen that no ranking is possible between the two constraints that make up the GPT. What is left to examine is its postulation of modularity, as well as the actual status of the constraints Minimal Attachment and Late Closure. Do they both play a role in parsing? If so, why is it so difficult to rank them? Is there another constraint at work in addition to Minimal Attachment and Late Closure? If they do not both influence parsing, does any one of them? To address these questions, we turn to more psycholinguistic results. In Section 2.2.1, we illustrate that the constraint set {Minimal Attachment, Late Closure} is not 'complete', in that there are other factors (lexical frequency information) involved in the parsing process. We also illustrate that the set {Minimal Attachment, Late Closure} is not 'sound', as the constraint 'Minimal Attachment' is seen to have a null effect on the parsing process. Section 2.2.2 will argue that some version of Late Closure must be involved in constraining the HSPM. Section 2.2.3 will illustrate that context and real-world knowledge do not influence initial parsing decisions. This will set the stage for the development of our OT system, for we will have enumerated the constraints that are

involved in parsing, as well as excluded constraints that have been claimed to be involved as such.

2.2 More Psycholinguistics

2.2.1 Role of Expectations

One of the most important papers for our work is TM. It works to serve two important functions. First, it demonstrates that Minimal Attachment is just a red herring and may (should) be dispensed with. Second, it introduces the importance of the parser's expectations of what is to come next.

Recall the PP attachment ambiguities (5), (6) (cf. 1.3), (11), (12) (cf. 2.1.1) and (19), (20) (cf. 2.1.3). We saw that no single ordering of Conservativity and Locality is able to capture the parsing preferences in each of these examples. TM argue that the difference in processing time between the NP-attachments and the VP-attachments in each of these examples is due not to syntactic principles, but rather to differences in readers' expectations in likelihood of semantic content.

Although their argument is directed at Minimal Attachment, we may take it as an attack on both Minimal Attachment and Late Closure. For recall that it was generally thought (at the time of writing of TM) that Minimal Attachment >> Late Closure. Hence, to demonstrate flaws in the GPT, it was sufficient to illustrate the vacuousness of its strongest principle. Because of its (hypothesised) inferior status, Late Closure was largely left out of the debate. However, we have seen that there is in fact no ranking between the two GPT constraints, and hence no 'superiority' of one over the other. Furthermore, VP-attachments are favoured by Minimal Attachment (Conservativity), and NP-attachments are favoured by Late Closure (Locality). As the NP-attachment is processed faster in one example ((19), (20)) whereas the VP-attachment is processed faster in the other ((5), (6)), neither Minimal Attachment nor Late Closure is able to account for the data on its own. Thus, we view TM's result as an attack on both Minimal Attachment and Late Closure. It will be found, however, that Late Closure has some independent evidence for its role in parsing, whereas Minimal Attachment does not.

Below, we present a brief overview of the reasoning found in TM that establishes their conclusions.

TM's starting point is the hypothesis that it is expectations of content rather than Minimal Attachment that leads to differences in processing time. To test this idea, they examine the sentence pairs from earlier studies that were used to help establish the robustness of Minimal Attachment.[12] The hypothesis is that the VP-attachments are actually closer to people's expectations than the NP-attachments, thus leading to more efficient processing time.[13] If the hypothesis is verified, the next step is to construct sentence pairs where the NP-attachments are rated closer to people's expectations than the corresponding VP-attachments. The hypothesis, of course, is that the NP-attachments should be processed faster than the VP-attachments, thereby negating Minimal Attachment and introducing a central role for people's expectations into parsing theory.

In validation of their 'hunch', TM find that the VP-attachments from earlier studies are indeed closer to people's expectations than the NP-attachments. An example of such sentence pairs is (5), (6) (cf. 1.3). The set of sentence pairs where NP-attachments are closer to expectations than VP-attachments includes the pair (19), (20) (cf. 2.1.3). When these sentences are presented to an independent group of readers, the NP-attachments are processed faster than the VP-attachments. This result violates Minimal Attachment, which predicts that VP-attachments are preferred due to the reduced number of nodes they require. Thus, we have seen that in all cases, it is people's expectations, rather than syntactic principles (Minimal Attachment, Late Closure) that have the greatest impact on reading times. TM make the following remark:

> *The most important finding in this experiment was a highly significant effect for subjects' expectations and a null effect for minimal attachment on reading times.* (TM, p.605)

[12] Recall that Minimal Attachment prefers VP-attachment over NP-attachment when faced with PP-attachment ambiguities.

What is needed now is a way to cash out the idea that people's expectations are making them take longer in processing rather than the well-established syntactic principles. What kinds of expectations are they? Where do they come from?

Carpenter, Miyake, and Just (1995, hf. CMJ) propose that these expectations are based on the frequency with which one has encountered certain concepts and syntactic structures. It seems as though people implicitly accumulate statistical information about the frequency with which they meet certain structures. When a certain (ambiguous) phrase is encountered, the parser prefers to select the most frequent candidate structure. More specifically:

> *Current research suggests that the frequency information associated with each verb's argument or thematic role structure is another important factor constraining the process of syntactic ambiguity resolution.* (CMJ, p.99)

To illustrate, consider the verbs 'remember' and 'claim'. Both can take NP complements, as in 'I remembered the idea', as well as sentence complements, as in 'I remembered the idea was bad'. Actuarial analyses of English have demonstrated that 'remember' tends to be followed more frequently by NP complements, whereas 'claim' is followed more frequently by sentence complements (CMJ). CMJ claim that readers do in fact use such tacit statistical information when processing ambiguous areas of a sentence. When people come across 'remembered', for example, they are expecting an NP. This expectation affects processing time in the obvious way. If the expectation is met, the processing time is quick. Otherwise, another structure needs to be selected, adding significantly to the processing time of the phrase.

2.2.2 Late Closure

[13] Expectations, which are based on relative frequencies, can be measured by having subjects write whole sentence completions for sentence fragments. For example, subjects can be asked to complete 'Kazashi believed', and the frequencies of completions determine measurements of expectations.

We saw above that TM's work negates Minimal Attachment and Late Closure. Such a result is unsatisfying in a conceptual way, for both constraints are derived from working memory considerations. It 'makes sense' to think that such considerations would have an influence on parsing strategies. However, empirical validation is needed to support the positing of any and all constraints. As far as we are aware, none really exists for Minimal Attachment.[14] However, as remarked earlier (cf. 2.2.1), Late Closure does have some independent evidence suggesting its use in language processing. This evidence reveals itself largely when lexical frequency information is controlled for. We use adverbial attachment and relative clause attachment data to illustrate the role of Late Closure in the parsing process.

Consider the following sentence:

(24) John said Bill died yesterday.

It has been observed (Gibson et al., 1996) that there is a strong preference to attach 'yesterday' to the more local clause. In fact, the preference is so strong that even changing the tense so that the lower clause is incompatible with the adverb has little effect on initial interpretation. The initial preference is to attach low, thereby causing processing difficulty (Gibson et al., 1996)[15]:

(25) (#) John said Bill will die yesterday.

This preference for low adverb attachment is also found in Spanish (Gibson et al., 1996):

[14] Well, it does have some descriptive power, as discussed earlier. However, in the face of a large set of counter-examples, it is not possible to maintain Minimal Attachment as a principled constraint in our theory of parsing.

[15] Intuitively, this does not boil down to lexical frequency information. I shall present evidence later in this section illustrating that such attachment preferences are independent of lexical frequencies. As a small experiment to test this, I asked six native English speakers (four Canadian, two American), about their interpretation of the following sentences. All attached the adverb low in both sentences.
(A) I met the man I fought yesterday.
(B) I fought the man I met yesterday.
If the attachment preference were lexical, then we would see a low attachment in one, and a high attachment in the other. This was not found to be the case.

(26) Juan dijo que Bill se murió (# morirá) ayer.
'John said Bill died (# will die) yesterday'.

Further evidence for Late Closure comes from relative clause attachment data. Consider the following sentence:

(27) The journalist interviewed the daughter of the colonel who had had the accident.

There is an observed preference to attach low, i.e. the preferred interpretation is that 'the colonel' had the accident, not the daughter (Gibson et al., 1996). There are many data supporting the claim that low relative clause attachment is preferred.[16]

Interestingly, this preference to attach low is *not* found in Spanish. Rather, the study carried out by Gibson et al. (1996) indicates that the preference is for the relative clause (RC) to attach high. For example, consider the Spanish translation of (27):

(28) El periodista entrevisto a la hija del coronel que tuvo el accidente.

The preferred interpretation is for 'la hija', and not 'del coronel', to have had the accident. Cuetos and Mitchell (1988) use these data to argue that Late Closure is not used in all languages, and in particular, is not used in Spanish.

However, because we are working in OT, we assume that all constraints are active in all languages. Thus, a more plausible explanation for the variance in data is that there is another constraint (not reducible to lexical frequency information) competing with Late Closure that prefers non-local attachments. For the sake of concreteness, let us call this hypothesised constraint 'Early Closure'. To explain the data, one simply posits that 'Early Closure' outranks "Late Closure' in Spanish, whereas in English the ranking is revered.

[16] See, for instance, Gibson et al. (1996), and Hemforth et al. (2000).

In order to test such a hypothesis, Gibson et al. (1996) modify the above experiment to obtain some surprising results.[17] Before discussing the details of their work, it is necessary to make a slight modification to our terminology. Following Gibson et al. (1996, p. 26), we replace the constraint 'Late Closure' with the more general 'Recency Preference':

(29) Recency Preference: Preferentially attach structures for incoming lexical items to structures built more recently.

The added generality comes from the fact that Late Closure chooses one attachment site over other alternatives, but does not rank them. Recency Preference, on the other hand, ranks all potential attachment sites; the more local the attachment site, the greater its rank.[18]

The above scenario involves RC attachment ambiguities where there are two possible NP attachment sites: NP_1 NP_2 RC. Gibson et al. (1996) perform a study of RC attachment preferences where there are three possible NP attachment sites: NP_1 NP_2 NP_3 RC. An example of such a sequence is the following:

(30) Pedro turned on [NP_1 the lamp near [NP_2 the painting of [NP_3 the house]]] [CP that was damaged in the flood]

To help determine preferences among the three NP sites, Gibson et al. (1996) disambiguate the possible attachments using number agreement so that only one NP site is grammatically available. They then measure reading times and gather on-line grammaticality judgements on the disambiguated versions of the RC attachments. For example, in the following, only NP_1 is available for attachment:

[17] There are actually two hypotheses being tested. First, that Late Closure *is* active in Spanish as well as English (indeed, it should be active universally). Second, that there is a second constraint, possibly related to what we have loosely dubbed 'Early Closure', that can account for the cross-linguistic differences in attachment preferences.

(31) [NP_1 la lámpara cerca de [NP_2 las pinturas de [NP_3 las casas]]][CP que fue dañada en la inundación]

'the lamp near the paintings of the houses that was damaged in the flood'

Assuming Recency Preference is the dominant factor in English, there is a clear prediction of the order of attachment preferences: attachment to NP_3 should be preferred over attachment to NP_2 which should be preferred over attachment to NP_1. We represent this notationally as NP_3 >> NP_2 >> NP_1. In Spanish, assuming 'Early Closure' is the dominant factor, the predicted order of attachment preferences is: NP_1 >>NP_2 >>NP_3. However, in both English and Spanish, in both grammaticality judgements and reading times, the observed preference is as follows: NP_3 >> NP_1 >> NP_2. The most preferred attachment site is the most local, followed by the least local, with the middle NP the least preferred of them all.

These results reveal important information about the nature of the constraints. First, as there is a local attachment preference in Spanish as well as in English, Recency Preference must be active in both languages. This should not be surprising, for we saw that Recency is also involved in adverb attachments in both English and Spanish. Second, the 'other' constraint must satisfy several conditions. First, it should be such that it does not affect attachments to verb phrases, for we saw that low attachment is preferred in adverbial attachment ambiguities in both English and Spanish, whereas high attachment is preferred in Spanish two-NP relative clause ambiguities. Second, it should be such that it can account for the fact that in both English and Spanish, the linear order of attachment preferences is not maintained when moving from two-NP to three-NP site ambiguities. In particular, it should be able to account for the fact that in Spanish, the most preferred attachment site goes from the highest NP to the lowest NP in the transition from two to three NP sites, while in English the least preferred attachment site goes from the highest NP to the middle NP.

[18] Note that Recency Preference can be very easily accommodated in OT.

Gibson et al. propose a constraint they call 'Predicate Proximity':

(32) Predicate Proximity: Attach as close as possible to the head of a predicate phrase.

The motivation for this constraint comes from their assumption that all grammatical utterances have a predicate at their core. If, as is the case with the HSPM, resources are limited, thereby restricting the number of attachment sites that can be left open, those sites associated with a predicate phrase will be more readily available than others, for they are absolutely essential for correctly interpreting grammatical phrases. Note that Predicate Proximity does not differentiate between multiple VP (predicate) attachment sites, so that adverbial attachment preferences are decided by Recency Preference. Thus, it satisfies one of the criteria required of the 'other constraint', as discussed in the preceding paragraph. Furthermore, in RC attachment ambiguities, Predicate Proximity favours high attachment, whereas Recency Preference favours low attachment. To account for the typological data, one simply posits that in Spanish, Predicate Proximity outranks Recency Preference, while in English the ranking is reversed. The explanation for cross-linguistic differences in the strength of Predicate Proximity is attributed to word order requirements imposed in the languages. For example, because English is (pretty strictly) SVO, the verb's arguments are never really 'far' from it, thus minimising the need for Predicate Proximity to be strong. Spanish, on the other hand, allows for more variable word order, such as VOS, thereby distancing the subject from the verb. This leads to a greater need for a strong Predicate Proximity constraint.[19] Thus, Predicate Proximity is seen to carry with it both descriptive and explanatory power. It is also predictive: languages with SVO or OVS word order should have a low-ranked Predicate Proximity constraint, whereas those with SOV, OSV, VOS, and VSO should place greater value on Predicate Proximity.

Unfortunately, as declared in Chapter 0, the constraint set {Predicate Proximity, Recency Preference} is unable to account for the parsing preferences in three-NP site ambiguities. Although it manages to predict preferences in two-NP relative clause ambiguities quite

[19] A 'strong' constraint can be thought of cognitively as a higher level of activation.

easily, it cannot predict the preference order NP_3 >> NP_1 >> NP_2 found in both English and Spanish. The best that this constraint set (as a standard OT system) can do is predict a linear preference order, i.e., either NP_1 >> NP_2 >> NP_3 or NP_3 >> NP_2 >> NP_1. In the weighted cost framework outlined in Gibson et al. (1996), however, the constraint set readily accommodates the data that it mishandles as an OT system. As we are searching for an OT analysis of the HSPM, this constraint set clearly won't suffice. In fact, as we remarked earlier, the constraint set {Predicate Proximity, Recency Preference} is at least 'incomplete', for it does not manage to predict parsing preferences arising due to lexical frequency information. This leads us to believe that there may be an alternative constraint set that is more reflective of the psychological constraints guiding the HSPM such that a translation of the constraints into a standard OT system would be able to predict the data discussed above. Additionally, we will illustrate later in the paper that the notion of Predicate Proximity is actually just a special case of another constraint involved in the functioning of the HSPM.[20] Thus, there is no need to introduce it as a free-standing constraint.

We introduce, instead, a pair of constraints that, in conjunction with lexical frequencies and Recency Preference, are argued to be sufficient in describing the HSPM. Before introducing the constraints, we ask the reader to allow a minor digression. An overriding feature of the system we are developing is that each of the proposed constraints contributes to the computational efficiency of the HSPM. Lexical frequencies do so by biasing the HSPM to a particular structure. In addition to reducing the computational complexity of the search for an optimal parse, this has the added effect of adding reliability to the parsing mechanism. In games of chance, where resources are scarce, it is safe to go with the most likely outcome. Recency Preference contributes to efficiency by reducing the load on working memory. By working as locally as possible, the HSPM reduces the number of concepts and structures required to be held and manipulated in working memory. The two constraints we will shortly introduce are similarly motivated by the notion of computational efficiency, which we take to be the most striking aspect of the HSPM.

[20] This constraint is a revised formulation of Recency Preference that will be introduced in Section 3.2.2.

The first constraint, which we call SALIENCE, uses a set of discourse entities ordered according to 'salience' to guide the HSPM in making anaphoric attachments:[21]

(33) SALIENCE: Anaphoric expressions should preferentially bind to the most salient sentence/discourse entity. As such, modifiers (eg. Relative clauses) that contain anaphoric expressions preferentially attach to the maximal projection of the most salient sentence/discourse entity.

What motivates the above constraint? First, it provides a heuristic that the HSPM can exploit to speed up the interpretation process. Rather than have to determine the best antecedent using an exhaustive search, the HSPM has a constraint that marks against attachments to non-salient entities. In effect, the HSPM is guided, or biased, towards salient items for purposes of anaphoric attachment. Second, relative clauses are 'anaphoric', in the sense that they refer to previously introduced discourse entities with the use of relative pronouns. In this regard, they differ from other kinds of attachments, such as PP attachments and adverbial attachments, which are not anaphoric. Thus, SALIENCE has nothing to say about adverbial attachments, thereby leaving Recency Preference to decide the fate of adverbial attachments. Hence, it satisfies one of the requirements outlined above that we argued must be met by the 'other' constraint.

However, note that this rests on a particular notion of what it means for a discourse entity to be 'salient'. The definition of salience we assume is derived from various standard ideas in the literature. We assume that the factors that contribute to a discourse entity's salience are the following:

- Animacy – We assume this is a binary feature. For some entity x, either [+x] or [-x] holds, with '[+x]' holding if and only if x is animate. Of course, the salience of some entity x increases if it is animate, and decreases otherwise.

[21] I would like to thank my thesis supervisor Henk Zeevat for not only bringing to my attention the importance (and possible relevance) of 'salience' as a constraint here, but also in sitting with me week after week, hours on end, discussing/debating/arguing/brainstorming these ideas with me to ensure we got to the bottom of things.

- Grammatical Obliqueness – Following the standard literature, we assume the following set is ordered according to salience: {subject, direct object, indirect object, adjunct}.
- Recency – We assume, according to the working memory literature (Baddeley, 1986), that more recently introduced discourse entities are more salient than ones introduced earlier. This is due to well-established exponential functions associated with the decay of information in working memory. Note that here, this notion of recency is based on the introduction of discourse entities into the on-going discourse. It is meant to contribute to the characterisation of 'salience'. Thus, although there are clear parallels between this version of recency and the more general version of 'Recency Preference' outlined above, this one is restricted in use to discourse entities, and thus will be involved in violation markings of NP-attachments only.

Working memory considerations give birth to the second constraint we require. According to Baddeley (1986), entities introduced into working memory decay exponentially with time. Thus, if one wants to manipulate them, or add further information to them, one should do so quickly and efficiently, before they decay. Furthermore, due to the incrementality of the HSPM (cf. Immediacy principle), and its preference to work locally, each new piece of information should, ideally, be contributing to the most recently introduced items in working memory. In particular, suppose that item x has been introduced, followed by item y which modifies x. Then, assuming z is the next item to be introduced, in the ideal case, z should be adding something to y, not x. We formalise this with the following constraint[22]:

(34) *MOD: Do not excessively modify any thing or event.

[22] There are also clear relations between this constraint, and constraints familiar from linguistic pragmatics, such as Horn's 'I-Principle'. We do not elaborate on any possible connections between these ideas here, but just note in passing that the idea of 'efficiency' in communication is quite robust, and is not limited to psycholinguistic theory.

Linguistically, this translates into a preference for a minimum number of adjuncts modifying any particular NP or VP. Hemforth et al. (2000) argue that the presence of multiple entities in the local environment (in working memory) results in interference effects that ultimately cause decreased activation. They argue that the interference does not cause much disturbance in the two-NP site ambiguities, but does in the three NP-site ambiguities. We extrapolate from these ideas, as well as from basic intuitions, that one adjunct modifying an NP or a VP (or any other node, for that matter) is easily accommodated by the HSPM, but more than one adjunct is dispreferred. Hence, following Smolensky (1997), we restate (34) as (35):

(35) *MOD: Do not modify any XP with more than one adjunct.

In Chapter 3, we will turn these constraints into concrete, optimality theoretic style constraints that incur an absolute number of marks for violations incurred. However, we are still not done our psycholinguistic examination. In the next section, we analyse two pragmatic factors that have been proposed as constraints on processing, viz., context and real-world knowledge. After this, we will be ready to formalise our system of constraints. To summarise, so far we have seen that lexical frequency information, working memory considerations (given form via Recency Preference and *MOD), and the salience of discourse entities are important factors in determining parsing preferences.

2.2.3 Pragmatics: Context and Real-World Knowledge

It has often been claimed, especially in the connectionist literature, that context and real-world knowledge are involved in the process of sentence comprehension. Certainly, the claim is quite plausible, for contextual and real-world knowledge effects run rampant through the cognitive science literature. However, Carpenter et al. (1995) inform us that pragmatic factors do not affect initial parsing preferences. For example, consider the past tense/past participle ambiguity in (36):

(36) The defendant examined by the lawyer shocked the jury.

The ambiguity in this sentence arises at the word 'examined', for it can be interpreted as either the main verb or a past participle. A garden-path effect is observed with this sentence, for the main verb interpretation is normally preferred. In sentence (37), the main verb interpretation is ruled out on semantic/pragmatic grounds due to the inanimate status of the head noun:

(37) The evidence examined by the lawyer shocked the jury.

However, a garden path effect is still observed with this sentence (Carpenter et al., 1995). The pragmatic knowledge is unable to override the main verb interpretation.

As another example, consider the garden path-sentence:

(38) The florist sent the flowers was pleased.

Rayner et al. (1983) illustrate that by making the sentence more plausible, i.e. more consistent with real-world knowledge, as in (39), the garden-path effect remains:

(39) The performer sent the flowers was pleased.

The available evidence suggests that the most important factor in these ambiguities is the frequency with which the verb in the given structure occurs as a past tense as opposed to a past participle (Carpenter et al. 1995, Gibson and Pearlmutter 1998). In the case of 'examined', for example, this form is more likely to occur as a past tense than a past participle, and hence the observed garden path effect associated with sentence (37), even though the interpretation is ruled out on pragmatic grounds.

The evidence therefore demonstrates that there is a null effect of pragmatic factors on *initial* parsing decisions. However, that is not to say that they do not have a role to play in parsing at all. Carpenter et al. (1995) indicate that context is found to influence parsing decisions only when lexical frequency information does not bias the parse in any

one direction. However, it must be emphasised that *initial* parsing decisions are not affected by pragmatic factors; they only seem to influence the HSPM once the other factors, viz., lexical frequencies, Recency Preference, *MOD and SALIENCE have been exhausted without a solution to the parsing problem.[23]

We are now in position to begin the formulation of our system of constraints. We have pursued an examination of the psycholinguistic literature to obtain an in-depth understanding of 'what is really going on' in the HSPM. Our study has determined that the following seem to be the constraints that guide the computations of the HSPM: lexical frequency information, working memory limitations, and a preference for anaphoric expressions to bind to salient discourse entities. Pragmatic factors are seen to influence parsing decisions only if the HSPM is unable to determine an optimal parse. Our goal is the formulation of an OT system of constraints that characterises the HSPM. In the next chapter, we translate lexical frequencies, SALIENCE, *MOD and Recency Preference into OT style constraints that give absolute measures of markedness. As pragmatic factors are not constituents of the HSPM, they will not play a role in the discussion to follow.

3. Parsing in OT

3.1 Preliminary Remarks

Recall from Section 2.2.1 that statistical information is used in the process of determining optimal parses. The results outlined in that section demonstrate the role of implicit learning in sentence comprehension. People are tacitly picking up on and using statistical information in processing linguistic inputs. This acquisition and use of statistical information suggests that a connectionist network is being employed at some point to carry out the task of parsing. This is a very intriguing prospect because humans don't learn most things with statistical learning. In particular, connectionist models

[23] Note that the notion of 'salience' makes crucial use of context, in the sense that entities are introduced into the context and referred to later using anaphoric expressions. However, we may think of this as a 'referential context', not to be confused with a 'propositional context' that is familiar from the literature. Nonetheless, it is an interesting question as to what the relation is between a referential context and a propositional context. We reserve commenting on this line of reasoning for future work.

traditionally perform quite poorly at language learning tasks (of any interesting level of complexity). The results of Section 2.2.1 illustrate that there may be a close connection between language and neural nets after all. Furthermore, they bring back a role for implicit, associative learning that was on the receiving end of Chomsky's critique all those many years ago.

Prince and Smolensky (1997) state that there is a close connection between optimality theory and connectionism. The important difference between them lies in the property of strict domination (McCarthy, 2002).[24] This is the point where our dispute with GB's analysis of parsing in OT begins. They argue that an OT system that obeys strict domination will be unable to capture the parsing data. In light of the remarks of the preceding paragraph, GB's conclusion seems to carry further plausibility and support.

In response, we offer the following remarks to justify our attempt at expanding the domain of standard OT to encompass the HSPM. First, the mapping between high level symbolic behaviour, exhibited most clearly by the language faculty, and low level network computation is neither clear nor well-understood. Thus, the fact that the computational principles guiding OT differ from those guiding connectionist networks is not necessarily a negative point against standard OT. Until the relation between high level and low level computation is made clearer, there is no *a priori* reason why the two should be characterised by the same computational principles.

Furthermore, the translation of the 'garden path theory of sentence processing' into OT that was carried out by GB should not even be expected to be able to account for the data, for we showed that the garden path theory is both 'unsound' and 'incomplete'. Additionally, the constraint set {Predicate Proximity, Recency Preference} is similarly 'incomplete'. As such, we may reason that GB's OT translations are unsuccessful not because of an inadequacy in standard OT, but rather because the parsing systems in GB are unfaithful representations of the HSPM. Having examined the psycholinguistic

[24] Connectionist networks tend to assign numerical weights to constraints, and allow for multiple violations of lower ranked constraints to outweigh a lesser number of violations of higher ranked constraints. This is precisely what strict domination prevents from happening.

investigations into the constraints guiding the behaviour of the HSPM, we are in a position to represent these constraints in a more accurate OT system. In fact, as outlined earlier, OT is particularly well-suited for such a task due to the inherent nature of 'optimality' found in parsing systems (cf. Chapter 0). Thus, we continue our quest to develop an OT constraint system that can characterise the HSPM. In Section 3.2, we present translations into OT style constraints of lexical frequencies, Recency Preference, *MOD and SALIENCE. The resulting constraint set is argued to be a highly accurate representation of the HSPM, and hence more likely to succeed as an OT system than the constraint sets presented in GB.

3.2 Presentation of the OT Constraint System

In order to develop an OT constraint system, we need to do two things. First, we need to translate the above psychological results into OT style constraints. To do this, we need a precise characterisation of how the proposed constraints are violated. Second, we need to motivate a ranking for the languages under consideration (English and Spanish). The second task is actually not so difficult. The data all suggest that frequency information stored in lexical entries is the most important factor in the determination of optimal parses (cf. 2.2.1, Trueswell et al. 1993, Boland 1997). Thus, any OT constraint(s) associated with such lexical frequency information will be ranked higher than constraints corresponding to SALIENCE, Recency Preference and *MOD. We argue that typology results from permutations of the latter three constraints. We begin the formulation of our system with lexical frequencies.

3.2.1 Lexical Frequency Information in OT

How are we to represent the lexical frequency information that is used by the HSPM? What kinds of constraints should be formulated to enable the HSPM to access lexical frequency information in the course of its computations? Modern psycholinguistics has done an admirable job of highlighting the importance of lexical frequency information to the HSPM. However, it has been remarkably silent on giving explicit formalisms describing, on the one hand, how the lexical frequency information is to be represented and, on the other hand, the mechanisms by which the HSPM accesses this information.

We use this silence as an opportunity to present an explicit formalism describing precisely these features. Doing so will require us to expand the OT picture of the language faculty to frontiers that have been heretofore avoided, namely relations of the language faculty to other cognitive faculties. In Section 3.2.1.1, we introduce a constraint called 'PROB-ACC' whose sole function is to access the probability information that is stored with lexical entries. We also outline architectural assumptions about the structure and organisation of the lexicon that allow PROB-ACC to carry out its computations in an efficient manner. Additionally, we discuss the interaction between PROB-ACC, the lexicon, and working memory. It is this latter kind of discussion, where aspects of the language faculty are situated in a general cognitive setting, that have, as far as I am aware, been largely neglected in the OT literature.

3.2.1.1 Architectural Assumptions

First, we assume that each lexical item has an ordered structure, where its subcategorization information is ordered according to the frequency with which it has been encountered. For example, consider the verbs 'remember' and 'claim'. Recall (cf. 2.2.1) that 'remember' is more likely to be followed by an NP complement, while 'claim' is more likely to be followed by a sentence complement. This information is assumed to represented in an ordered set as follows: 'remember': {< __ NP>, < __ S>}, 'claim': {<__ S>, <__ NP>}.

Second, we assume that verbs have another ordered structure associated with them, viz., probability information describing the frequency with which they are used in the present/past tense as well as the frequency with which they are used as a present/past participle. This information is assumed to be represented as an ordered set, much like above. For some verb V, suppose that it is most frequently used in the present tense, then as a past participle, then in the past tense, and then as a present participle. We assume this information is represented as follows: 'V': {Present Tense, Past Participle, Past Tense, Present Participle}.

Third, we assume that the HSPM has communicative links with both working memory and the lexicon. In particular, we propose a constraint called 'PROB-ACC' that serves this communicative function. Suppose that the current parse after 'k-1' words is X*, and that the current word in the input is W_k. We assume that this information is stored in working memory, and that PROB-ACC accesses this information, so that in a sense it 'knows' what the current word and most recent parse are. In evaluating candidate structures, PROB-ACC uses the ordered lexical information to help determine the most optimal path the parse should take.

3.2.1.2 PROB-ACC Explained

Here we formalise the method by which 'PROB-ACC' works in unison with working memory and the lexicon to help determine optimality. We assume the same phrase structure grammar as before (cf. Section 2.1). Let X* be the parsed structure of input $(W_1,\ldots,W_{k-1})$. At W_k, we assume that held in working memory is the information that X* is the most recent parse, and that W_k is the most recent word. At W_k, GEN will output a set of candidates. Let $\{X_1, X_2,\ldots\}$ be the output set.[25] Let X_i be an arbitrary member of this set, so that it is a candidate parse of the new input. During evaluation of X_i, PROB-ACC will make use of different sets of lexical information depending on the type of ambiguity the HSPM is faced with. From observing the data, there seem to be three prominent types of ambiguities the HSPM encounters.

One type of ambiguity the HSPM is commonly faced with is deciding which category the current word W_k is a constituent of when it follows a verb W_{k-1}. We call this a Type-1 ambiguity. For example, consider the sentence 'The professor noticed the student was not paying attention'. When the HSPM gets to the word 'the', it needs to decide whether 'the' belongs to a simple NP or whether it is part of a CP argument of 'noticed'. In this type of ambiguity, the expectations of what argument/adjunct the verb will take are the main determinant of the path the parse should take. Essentially, the HSPM must *select* the *best* (i.e. most probable) argument/adjunct for the verb W_{k-1}. To help the HSPM

[25] This will generally be a small finite set, for only grammatical candidates are considered.

make its selection, PROB-ACC utilises the ordered subcategorization information of the word W_{k-1}, viz.,

(40) 'W_{k-1}: {<__ y_1>,… <__ y_n>}, where the y_i are the phrase categories licensed by W_{k-1}.

To simplify our notation, we represent this information in the manner that gives rise to the order itself, viz., as numerical probability values. Hence, we represent (40) as the following set: {$\pi(W_{k-1})(y_1)$,…$\pi(W_{k-1})(y_n)$}, where $\pi(W_{k-1})(y_i)$ is the probability that W_{k-1} takes the category y_i. Applying this to our example, PROB-ACC uses the values from the set {π(noticed)(NP), π(noticed)(CP)} to help guide its parse. Whichever of the two values is greater will cause a preference for the HSPM to parse the input so as to match the structure associated with the higher value. We will cash this out more formally shortly, after presenting the other types of ambiguities faced by the HSPM.

A second type of ambiguity faced by the HSPM occurs when it is known which category the current word W_k belongs to, but it needs to be determined where this category should attach. Call this a Type-2 ambiguity. As an example, consider the following sentence, taken from TM:

(41) The thieves stole all the paintings in the museum.

The point of ambiguity occurs at the word 'in'. Here, one knows that the word 'in' belongs to PP, but the relevant question is whether the PP should attach to the VP or the NP. To help decide this, the subcategorization information that is accessed is the following: {π(steal)(PP), π(painting)(PP)}. In this case, π(painting)(PP) >> π(steal)(PP) (TM). This translates into a preference for the PP to attach to the maximal projection of 'painting'.

The third type of ambiguity faced by the HSPM involves the verbal morphology of the current word. For example, in the sentence 'The horse raced past the barn fell', there is

an ambiguity at the word 'raced': Is it to be interpreted as the main verb or as part of a reduced relative? In order to determine this, PROB-ACC must access the frequency with which 'raced' occurs as a past participle and the frequency with which it occurs in the past tense. This information can be represented (using the π-notation as above) as π(raced)(Past Tense) and π(raced)(Past Participle). Whichever of the values is higher will push the parse in its direction. Note that this ambiguity only occurs when the past participle and past tense forms of a verb have the same morphological marker. For example, this type of ambiguity occurs with 'raced', but not with 'drive'.

Let us return to the issue of formalising the constraint PROB-ACC. If, at the current word W_k, the HSPM is faced with a Type-1 ambiguity, PROB-ACC will access from the lexicon information from the set $\{\pi(W_{k-1})(y_1),\ldots,\pi(W_{k-1})(y_n)\}$. Let us assume this set is ordered such that $\pi(W_{k-1})(y_i)$ is higher in the order (>>) than $\pi(W_{k-1})(y_j)$ only if $i < j$. Now, let X_i be a candidate output. Suppose that X_i parses the current word W_k so that it belongs to category y_j. Thus, the structure is $W_{k-1}(y_j)$. In its EVALuation of X_i, PROB-ACC simply observes the position of $\pi(W_{k-1})(y_j)$ in the ordered set $\{\pi(W_{k-1})(y_1),\ldots,\pi(W_{k-1})(y_n)\}$, and assigns it j-1 violations. Thus, if j=1, i.e. $W_{k-1}(y_j)$ is the most probable parse, the candidate structure X_i receives no violations. If j=2, the candidate structure receives one violation. If j=n, i.e. the candidate structure is the least likely parse, it incurs n-1 violations.

If the HSPM is faced with a Type-2 ambiguity at the word W_k, it needs to determine where the category to which W_k belongs (y_j, say) should attach. Let X_i be a candidate output such that y_j attaches to the maximal projection of word W_r, where W_r is a word that precedes W_k in the current phrase allowing y_j attachment. The constraint PROB-ACC accesses from the lexicon a subset of $\{\pi(W_1)(y_j),\ldots,\pi(W_{k-1})(y_j)\}$, where $W_1,\ldots,W_{k-1}$ are the words preceding W_k in the input string. The subset will consist of those elements that allow y_j as a possible attachment. If the position of $\pi(W_r)(y_j)$ in the order is 's', say, then X_i incurs s-1 violations of PROB-ACC.

If the HSPM is faced with a Type-3 ambiguity at the word W_k, PROB-ACC accesses from the lexicon a subset of the information $\{\pi(W_k)(T_1), \pi(W_k)(T_2), \pi(W_k)(T_3), \pi(W_k)(T_4)\}$, where the T_i are permutations of the past/present tense and past/present participle forms of the verb W_k. The members of the subset are those elements of the above set that allow the same overt form of the verb with different T_i's. Let us assume that the given subset is ordered by frequency. In EVALuating a candidate output X_i, where X_i parses W_k such that it carries T_j, if the position of $\pi(W_k)(T_j)$ in the ordered subset is 'r', say, then X_i incurs r-1 violations of PROB-ACC.

3.2.2 Recency Preference

In the literature on ambiguity resolution, formulations of a 'recency preference' have been motivated by working memory considerations, as well as by constraints on interpretation.[26] For example, Node Locality/Late Closure was formulated largely so as to minimise the expenditure of computational resources by working memory. Recency Preference, as formulated by Gibson et al. (1996), was also guided by considerations of working memory. Gibson and Pearlmutter (1998) further suggest that in unambiguous sentence processing, greater distance between heads and their dependents leads to greater processing difficulty. This is, again, rooted in considerations of working memory and interpretability.

We would like to suggest that the latter notion of head-dependent distance is sufficient as a constraint defining 'recency preference'. No further formulation of recency is necessary. We think this is very plausible, for it captures at once the attempt to minimise working memory resources, as well as interpretation requirements, where it is well known that the closer that heads and dependents are to one another in a given structure, the easier it is to interpret that structure. This also provides a motivation for Gibson et al. (1996)'s 'Predicate Proximity' constraint, for the latter is just a special case of the constraint we are proposing now.

In the spirit of OT, because the constraint works to minimise head-dependent distance, we name the constraint *H-D. In (42), we provide a definition of the constraint *H-D:

(42) *H-D: Minimise distance between heads and their dependents.

Of course, that won't do as an OT constraint, for we need to assign an associated markedness calculus. We propose the following:[27]

(42') A structure Y receives n violations of *H-D if there are n maximal projections intervening between the maximal projection of the head H and the maximal projection of the dependent D.

An example is surely in order. Consider sentence (24), reproduced below as (43):

(43) John said Bill died yesterday.

In the case where 'yesterday' attaches to the VP headed by 'said', the structure incurs four violations of *H-D, one each for the intervening CP, IP, NP and VP. In the case where 'yesterday' attaches to the VP headed by 'died', *H-D incurs no violations at all, for there are no maximal projections intervening between the maximal projection of 'died' and the maximal projection of 'yesterday'. Thus, *H-D prefers the more local attachment, and therefore guides the HSPM to local attachment, much as Late Closure and Node Locality and Recency Preference had done before.

3.2.3 Salience

We conceive of SALIENCE as a discourse-entity based constraint. Maximal projections of highly salient entities are preferred for attachment decisions. We assume that many factors contribute to an item's salience, as outlined in Section 2.2.2. We recall for the

[26] Namely, the ease of interpretation of structure xAy, where 'y' attaches to 'x', varies inversely with the size of 'A'. Of course, this constraint on interpretation is (probably) reducible to considerations of working memory, but we are happy to include the interpretation constraint even if we are being redundant.

reader that the factors contributing to an item's salience are animacy, recency, and grammatical obliqueness. What remains to be formulated is an explication of how violations of SALIENCE are to be calculated. In our system, violations of SALIENCE are calculated by summing the violations incurred by a candidate structure of each of the factors contributing to SALIENCE. We now express how such violations are distributed.

Suppose that in a candidate structure X, some anaphoric expression attaches to a discourse entity x. 'Animacy' can contribute at most one 'SALIENCE' violation: either x is animate or it isn't. In the case that it is, there are no violations of SALIENCE. In the case that it isn't, there is one violation of SALIENCE. 'Grammatical obliqueness' contributes violations to a candidate structure X as follows. If entity x plays the role of subject, X incurs no violations of SALIENCE. If x is a direct object, then X receives one violation. If x is an indirect object, X incurs two violations of SALIENCE, and in case x is an adjunct, X receives three SALIENCE violations. Finally, 'recency' contributes to SALIENCE violations as follows. If the NP headed by x is the most recent possible attachment site, X incurs no violations. The number of violations increases linearly with the non-locality of the possible attachment sites, i.e. one violation per allowable NP that is passed over in making attachments.

Before translating *MOD into an OT constraint, we should provide some further remarks about the 'recency' that contributes to SALIENCE and the 'recency' that contributes to *H-D. There is a possibility that a candidate structure may receive 'recency' type violations from both SALIENCE and *H-D. It is up to us now to argue that this is not problematic. And it isn't, for the fact that it satisfies the following important meta-constraint we may impose on OT constraint systems: There should be a one-one mapping between constraints and the things they are marking. Problematic constraint systems arise if this relation is many-one, for such a constraint system results in redundant and vacuous use of constraints, and if the relation is one-many, which results in a constraint system deplete of typology. We will illustrate later in this text (Chapter 4) that our

[27] This notion of distance is somewhat arbitrary. Another possible option is 'number of intervening words' or, even better, 'number of intervening content words'.

system indeed gives rise to typological variance in the data. In the next paragraph, we illustrate that the system avoids redundancy, hence allowing our constraint system to satisfy the meta-constraint on constraint systems, hence allowing our system to satisfy an important adequacy criterion.

Observe that the two constraints SALIENCE and *H-D are marking for two different things. The former is geared towards ensuring that attachments are made to 'salient' discourse entities. From the working memory literature (Baddeley, 1986), we have learned that an item's salience varies with time/distance. Thus, the notion of salience happens to encompass a notion of recency, but this notion is only relevant in as much as it contributes to the grander notion of salience. In the constraint *H-D, we are attempting to minimise the distance between heads and dependents. This relation is a structural one ranging over all head-dependent relations. The latter may be formulated in purely syntactic terms, whereas the former may be formulated without any mention of syntax whatsoever. Each constraint is marking for different things: one is marking for salience, the other is marking for head-dependent distance. It is this fact that saves our constraint system from vacuous, redundant use of 'recency'.

3.2.4 *MOD

*MOD is a constraint that marks against excessive use of modifiers, where 'excessive' means more than one. Violations of this constraint are calculated as follows. Let X be a candidate structure. Take each node XP in X that is modified by more than one adjunct. For each such node, if it is modified by n adjuncts, then it contributes n-1 violations of *MOD. The total number of violations incurred by X is the sum of all such values. This is perhaps one of the few cases where an algorithm (high level) can explain the method better than words. In the following, suppose X contains nodes $XP_1,\ldots,XP_m$, let 'num(XP)' be the number of adjuncts modifying XP, and let 'N' be the number of violations X incurs of *MOD..

Set N=0

for j=1,…,m

if num(XP_j) >1, then do

$N \leftarrow N + (num(XP_j)-1)$

end if

end for

Output N

4 Testing the Theory

4.1 Preliminary Remarks

In our investigations, we have come across five common types of structural ambiguities. We argue that these five can all be reduced to just three, viz., 'Type-1', 'Type-2' and 'Type-3' ambiguities. Below we enumerate the five ambiguity types, and illustrate how they are really just instances of the three types of ambiguities outlined in Section 3.2.1.2.

(A) CP-continuations vs. NP-continuations (of a verb)

In these ambiguities, the HSPM must select the appropriate subcategorization frame of the preceding verb. In particular, although it is known that the current word is a noun, should it be interpreted as part of a CP argument of the verb or as an NP argument of the verb? This is, of course, just a Type-1 ambiguity (cf. 3.2.1.2). An example of such a sentence is 'Ashesh knew Randeep liked Samira'. The ambiguity occurs at the word 'Randeep'.

(B) Reduced Relative vs. Main Verb

Here, the ambiguity is resolved by determining whether the verb (the current word) should be parsed as part of a reduced relative or as a main verb. This is a Type-3 ambiguity. An example of such an ambiguity is the ever-popular 'The horse raced past the barn fell', with the ambiguity occurring at the word 'raced'.

(C) NP-Attachment vs. VP-Attachment (PP attachment ambiguity)

In these ambiguities, the HSPM must determine whether the current word is an argument/adjunct of the verb or an adjunct modifying the noun. For example, consider the sentence 'I read the article in the magazine'. At the word 'in', should the parse be

such that ‘in’ modifies ‘article’ or such that it is an argument of ‘read’? This is, of course, nothing more than a Type-2 ambiguity.

(D) Adverbial Attachments

When presented with an adverb that has multiple possible attachment sites, the HSPM must determine which of them to attach to. This is an instance of a Type-2 ambiguity. An example of such an ambiguity is found in the sentence ‘John said Bill died yesterday’.

(E) Relative Clause (hf. RC) attachments

These ambiguities occur when the HSPM is presented with an RC that has multiple possible NP attachment sites. This presents the HSPM with a Type-2 ambiguity. An example of such an ambiguity is found in the sentence ‘The journalist interviewed the daughter of the colonel who had the accident’. The RC could modify either ‘the daughter’ or ‘the colonel’, and it is the karma of the HSPM to have to figure out which one is the case.

The above enumeration indicates that many of the structural ambiguities found in the literature are actually just instances of Type-1, Type-2, and Type-3 ambiguities. This makes the job of the HSPM a whole lot easier, for it has to deal with a smaller contingent of ambiguity types. This is also scientifically elegant, for ambiguities can now be thought of more generally, without having to worry about construction specific details of each ambiguity. A further simplification occurs because the various kinds of structural ambiguities are now reduced to ambiguities in lexical information. The constraint PROB-ACC will directly access frequency information stored in the lexicon in order to help disambiguate the input. It is presented with a small ordered information set from the lexicon, and uses that information in the evaluation of candidate structures. Type-1 and Type-2 ambiguities present sets of subcategorization frames that compete for selection. Type-3 ambiguities present sets of lexical forms that differ in tense/aspect but share the same morphological form. The rank in the order of the lexical information exhibited by each candidate determines that candidate’s well-formedness with respect to PROB-ACC. In the event that PROB-ACC is unable to determine optimality from the lexical

information it is presented with, the other three constraints go to work to help determine the optimal parse of the given input.[28,29] This is all quite lovely, for it illustrates how the computation of an optimal parse can be reduced to a selection procedure at each step, which really justifies the marriage we have proposed between OT and the HSPM. The selection is based on ordered information in the lexicon, which is built up implicitly over the course of an individual's lifetime. The order is determined by the frequencies with which certain structures have been met. To repeat: the architecture outlined above ties together a seemingly disconnected array of structural ambiguities, reducing them to sets of lexical ambiguities cashed out as Type-1, Type-2 and Type-3 ambiguities. The constraint PROB-ACC is designed to explicitly capture the frequencies stored in the lexicon to help disambiguate inputs. In the event that it is unable to do so, for example if lexical probabilities are equal, then we have three constraints that capture what seems to be going on at the psychological level, viz., SALIENCE, *MOD and *H-D. We claim that this set of constraints is sufficient for explaining parsing preferences. It is the goal of the next section to illustrate that this claim is well-justified.

[28] Note that we do not insist that the order be total. For example, suppose verb V subcategorizes for NP's and for CP's. Suppose the frequency information stored in the lexicon is such that $\pi(V)(NP) = 0.52$, and $\pi(V)(CP) = 0.48$. In this case, it seems unlikely that PROB-ACC should distinguish between the two. Thus, it becomes the goal of psycholinguistic theory to determine the conditions under which two values satisfy the probabilistic order relation. For example, one can posit that '$\pi(V)(NP) >> \pi(V)(CP)$' iff $\pi(V)(NP) - \pi(V)(CP) > 0.1$, or something like that. What is important is that in a particular theory, there *should* be some such condition, such that it can be tested and put up to being falsified.

[29] Reinhard Blutner has pointed out to me the following problem with this formulation of PROB-ACC. Suppose, for the sake of exposition, that there are two candidate structures X and Y, and that the lexical frequency information is such that $\pi(X) >> \pi(Y)$. As per Footnote 28, suppose this means that $\pi(X) - \pi(Y) > r$, where 'r' is some numerical value between 0 and 1. Then PROB-ACC assigns the same number of constraint violations to structures X and Y regardless of the values of $\pi(X)$ and $\pi(Y)$, so long as $\pi(X) - \pi(Y) > r$. So, if 'r' is 0.1, then X and Y will incur the same distribution of constraint violations if $\pi(X) - \pi(Y) = 0.11$ and if $\pi(X) - \pi(Y) = 0.9$. This poses, a priori, a potential problem for the system. There are not enough data currently available (that I am aware of) that could be used to determine 'grades' of preferences that vary with the amount of difference between probability values. However, if it were determined that there is a mapping between preference strengths and probability differences, this could be easily accommodated in our system. Currently, we have the lexical frequency information is ordered as $\{X_1,\ldots,X_k\}$. To account for 'grades of preferences', the EVALuation component is altered as follows. Structure X_1 incurs no violations of PROB-ACC. For all other structures, if $\pi(X_1) - \pi(X_j) > nr$, where 'n' is an integer, then X_j incurs n PROB-ACC violations. More generally, for any two structures X_i and X_j, if $X_i - X_j > mr$, then X_j incurs m more violations of PROB-ACC than X_i. We await further psycholinguistic evidence to determine how, if at all, differences in probability values impact the strength of the preference. Doing so will determine the value of 'r' in the above formulation. For now, we maintain the null hypothesis that it is solely the order that is relevant in the assignment of constraint violations.

I think that should suffice for the preliminary remarks. In the following section, I shall, after having made the reader wait longer than I perhaps should have, test the theory on experimental data. We have four constraints in our system: PROB-ACC, *H-D, *MOD and SALIENCE. It is argued that in all languages, PROB-ACC is ranked higher than each of the other two constraints. This is because of the psychological finding that lexical probability information is the most influential factor in parsing decisions. Typology results from permutations of the constraints *MOD, *H-D and SALIENCE. In English, we have the following ranking: PROB-ACC >> *MOD >> SALIENCE >> *H-D. In Spanish, the ranking is: PROB-ACC >> SALIENCE >> *MOD >> *H-D. To simplify our analysis of the data, we will break up the analysis of examples into two cases: those where PROB-ACC differentiates among competing structures and those where PROB-ACC doesn't. That way, in each case, we can focus on the *relevant* constraints that decide among competing structures without getting bogged down by details that will be irrelevant to the determination of optimality. It is hoped that the reader will forgive the exchange of thoroughness of detail in favour of explanatory clarity. With the apologies out of the way, to the data at last.

4.2 Data Coverage

4.2.1 Where PROB-ACC is Enough

In this section, we illustrate how PROB-ACC accounts for various observed parsing preferences in the resolution of Type-1, Type-2 and Type-3 ambiguities. We begin with examples of Type-1 ambiguities.

4.2.1.1 Type-1 Ambiguities

Consider the following sentence:

(44) The student forgot the solution was in the back of the book.

The ambiguity occurs after the word 'forgot', for it can take either an NP or a CP. It is Type-1 because the HSPM must decide whether 'the' belongs to an NP argument of 'forgot' or a CP argument of 'forgot'. To help decide, PROB-ACC accesses from the

lexicon the information that π(forget)(NP) >> π(forget)(CP) (Trueswell et al., 1993). Thus, the NP argument reading will receive less violations of PROB-ACC than the CP argument reading. Hence, our system predicts the observed preference (Trueswell et al., 1993) for the NP continuation, and hence also the observed garden path effect associated with (44).

Note that what is important here is the *order* of the lexical probability information. We know that π(forget)(NP) is higher in the order than π(forget)(CP), and thus it will receive less violations or PROB-ACC. However, for concreteness, let us assume that NP and CP are the only possible arguments of the verb 'forgot'. Then, the parse that follows the NP argument reading will receive no PROB-ACC violations, for it is highest in the lexical order. The parse that follows the CP reading will receive one violation of PROB-ACC, for it is second in the order. Recall (cf. 3.2.1.2) that a parse that follows the nth ranked structure in the lexical probability order receives n-1 violations of PROB-ACC. The structure with the least number of PROB-ACC violations (if indeed there is such a structure) is judged optimal. In the following examples, it will suffice to illustrate that the lexical order dictates the observed preference. However, it is important to note that this is formally cashed out in terms of discrete numbers of constraint violations.

Consider now the following Type-1 ambiguity:

(45) I remembered the idea was bad.

The ambiguity here arises after the word 'remembered', which can take either a simple NP or a CP. PROB-ACC accesses from the lexicon the probability values π(remember)(NP) and π(remember)(CP). The former value is higher in the order than the latter (Carpenter et al. 1995). Thus, the structure [IP[NP I][VP[V remembered][NP the]]] incurs less violations than [IP[NP I][VP[V remembered][CP[C e][IP[NP the]]]]]. Therefore, the system predicts that there should be a garden path effect associated with (45), which indeed there is (Carpenter et al., 1995, Trueswell et al., 1993).

Consider now our final example of a Type-1 ambiguity, taken from Trueswell et al. (1993):

(46) The student realised the language was spoken in only one province.

Note that 'realise' subcategorizes for NP's as well, as in 'The student realised his goals'. In this case, π(realise)(CP) >> π(realise)(NP) (Trueswell et al., 1993). Thus, at 'the', the point of ambiguity, PROB-ACC will assign greater constraint violations to the NP-continuation than to the CP continuation, therefore predicting that the CP-continuation should be preferred. It is indeed found that there is no garden path effect associated with the sentence (Trueswell et al., 1993). Thus, the system captures the observed parsing preference in this example as well. Note that this example differs from the first two in that the CP-continuation is more likely than the NP-continuation. Note further that on the basis of a constraint like Minimal Attachment, this datum could not be accounted for, as the CP-continuation suffers more violations of Minimal Attachment (cf. Node Conservativity) than the NP-continuation. Note further that Late Closure (cf. Node Locality) also does not play any role in the parsing preference here. The system developed in our work captures precisely the psycholinguistic fact that frequency information associated with subcategorization information is highly influential in parsing strategies. This factor is missing from the analysis proposed by Gibson and Broihier (1998), which is one of the reasons why their system of constraints is unable to account for some very important data.

4.2.1.2 Type-2 Ambiguities[30]

Consider the sentence (47):

(47) Hans cut the apple with a blemish.

This presents the HSPM with a Type-2 ambiguity because it needs to decide where to attach the PP 'with a blemish'. At the word 'with', an attachment can be made to either

the VP headed by 'cut' or the NP headed by 'apple'. The parse at this point is [IP[NP Hans][VP[V cut][NP[D the][N apple]]]]. PROB-ACC accesses from the lexicon the information that π(cut)(PP) >> π(apple)(PP)[31]. Thus, the structure corresponding to the NP-attachment [IP[NP Hans][VP[V cut][NP[D the][N'[N apple][PP with]]]]] will incur more violations of PROB-ACC than the structure corresponding to the VP-attachment [IP[NP Hans][VP[V cut][NP[D the][N apple]][PP with]]]. Thus, a prediction of our theory is that (47) should take a longer time to process than, say, 'Hans cut the apple with a knife', where the PP attaches to the VP rather than the NP. This is indeed found to be the case, and so the system captures the observed data.

Next, consider the sentence:

(48) The bully hit the girl with a book with a bat.

The HSPM is faced with a Type-2 ambiguity at the first preposition 'with'. The PP can be either the beginning of a PP-argument or the beginning of a PP adjunct modifying the NP argument. The latter is correct, but the system should predict a preference for the former interpretation in order to predict the garden path effect associated with this sentence. From the lexicon, PROB-ACC learns that π(hit)(PP) >> π(girl)(PP), and so the EVAL component of the grammar distributes less violations of PROB-ACC to the VP-attachment than to the NP-attachment. Thus, the HSPM prefers the VP-attachment at this point in the parse, and suffers later on down the road, having misparsed the input. This prediction matches the experimental data, as a garden path effect is observed with (48).

In (49), we see another Type-2 ambiguity at the preposition 'in':

---

[30]All of the examples and data in this section come from Taraban and McClelland (1988).

[31] It might well be that the frequency information, like π(cut)(PP), depends on the actual preposition under consideration, so that it might be more correct to use π(cut)(PP_{with}), and π(cut)(PP_{in}), etc. However, we have no data to suggest that this is so (or not so). Thus, we proceed with the minimal assumption that the frequency with which the verb or noun takes a PP is the sole information that is used by the HSPM. This is actually preferred to the preposition-specific case, for this prevents our lexicon from becoming too 'bloated'. In any event, it is seen that we are able to capture the parsing preferences observed with PP-attachments with the minimal assumptions being made. Good.

(49) I read the article in the bathtub.

The parse at the point of ambiguity is [IP[NP I][VP[V read][NP[D the][N article]]]]. Again, the PP can either attach to the NP headed by 'article' or to the VP headed by 'read'. To help resolve this stalemate, PROB-ACC accesses the information that π(article)(PP) >> π(read)(PP) from the lexicon. Thus, the NP-attachment [IP[NP I][VP[V read][NP[D the][N'[N article][PP in]]]]] will incur less violations of PROB-ACC than the VP-attachment [IP[NP I][VP[V read][NP[D the][N article]][PP in]]]. Therefore, the system predicts that there should be a garden path effect associated with this sentence, as opposed to a sentence like 'I read the article in the magazine'. And, as you may have guessed, there is indeed found to be such an effect in the processing of (49).

Note that this result cannot be accounted for by Minimal Attachment (cf. Node Conservativity), for the preferred structure (the NP-attachment) incurs a greater number of Conservativity violations than the VP-attachment. However, this result is consistent with Late Closure (cf. Node Locality). On the other hand, Locality does not account for the parsing preferences observed in examples (47) and (48). Again, it is only by appealing to lexical frequency information that we can systematically account for each of the above parsing preferences.

As our final example of a Type-2 ambiguity that is handled by PROB-ACC, consider sentence (50):

(50) The thieves stole all the paintings in the night.

The story should be quite familiar by now. The HSPM is faced with an ambiguity at the word 'in'. Should this be interpreted as an NP-attachment or a VP-attachment? PROB-ACC accesses the fact that π(painting)(PP) >> π(steal)(PP) from the lexicon. Thus, candidate structures that interpret 'in' as a VP-attachment will suffer more violations of PROB-ACC than candidate structures interpreting 'in' as an NP-attachment. Thus, the

system predicts that there should be a garden path effect associated with this sentence (compared with, say, 'The thieves stole all the paintings in the museum'). This has been experimentally determined to be the case. Thus, the constraint system is able to capture the observed data.

4.2.1.3 Type-3 Ambiguities

In this section we look at ambiguities involving past tense/past participle interpretations of a verb. The verb 'examine' has 'examined' as both its past tense form, as in (51):

(51) The lawyer examined the evidence,

and as its past participle form, as in (52):

(52) The defendant examined by the lawyer shocked the jury.

It is found that 'examined' occurs more frequently as a simple past tense than as a past participle (Trueswell at al., 1994). Thus, the theory predicts that there should be a garden path effect associated with the processing of the by-phrase in (52)[32]. Furthermore, because lexical frequency information is more influential than other factors in initial parsing decisions, such as context and plausibility, forcing the past participle reading (by context or plausibility) should not alter the fact that a garden path effect is associated with the past participle usage of 'examined', as in (53)[33]:

(53) The evidence examined by the lawyer shocked the jury.

It is found that there is a garden path effect associated with the processing of the by-phrase in both (52) and (53) (Carpenter et al., 1995).

[32] This can be measured against unambiguous versions of the sentences, such as 'The defendant that was examined by the lawyer shocked the jury'.
[33] However, doing so may lessen the 'degree' of processing difficulty. For a lucid discussion of the notion of 'degree' of difficulty, or garden-path effect, see Trueswell (1996).

In related work, Trueswell (1996) confirms the hypothesis that the frequency of a verb's past participle form influences its processing complexity: The higher the frequency, the lower the processing complexity. Trueswell (1996) performs experiments where the initial noun can either be a good agent of the verb (as 'the lawyer' is for 'examined') or a bad agent (as 'the evidence' is for 'examined'). The study finds that in each case, the processing complexity of the reduced relative varies inversely with the frequency of its verb's use as a past participle. For example, 'searched' occurs as a past tense far more frequently than it does as a past participle. Conversely, 'accused' occurs more frequently as a past participle than as a past tense. It is found that when both verbs are placed in reduced relative clauses, such as in the sentences (54) to (57), the by-phrase disambiguating each sentence is processed with greater complexity in the case of 'searched' than it is in the case of 'accused':

(54) The room (that was) searched by the police contained the missing weapon (bad agent).

(55) The thief (who was) searched by the police had the missing weapon (good agent).

(56) The suspect (that was) accused by the investigator had no real alibi (bad agent).

(57) The witness (who was) accused by the investigator had no real alibi (good agent).

Unfortunately, Trueswell (1996) does not compare the reading times of reduced relatives versus main verb interpretations of verbs that have high frequencies of past participle occurrence. What these results show is that high participle frequencies make it much easier to process reduced relatives that contain those verbs as past participles. What the results *do not* show is that if a particular verb has a higher past participle frequency than a past tense frequency, then it should be easier to process sentences in which the verb occurs as a past participle (eg. Reduced relatives) than sentences in which it appears as a simple past tense (eg. Main clause). This remains an open question, but the results of Trueswell (1996) provide initial support for the conjecture that if π(V)(past participle) >>

π(V)(past tense), then the complexity of processing a reduced relative containing the verb V will be less than the complexity of processing a structure where V is the main verb.

4.2.2 Where PROB-ACC Does Not Determine Optimality

In this section, we will examine examples of sentences where PROB-ACC does not (as far as we are aware) resolve ambiguities on its own. We assume the impotence of PROB-ACC for two reasons. First, there was no available data (that we could find) by which we could claim that there is a frequency bias in the given examples. Second, in each case, it does not seem like there would be (a priori) any subcategorization frequency bias toward any particular interpretation. Thus, we assume that PROB-ACC does not differentiate between the candidates we will present in this section. In other words, we assume that all the candidates to be investigated are tied with respect to PROB-ACC. Thus, we will focus on their well-formedness with respect to the constraints *MOD, *H-D and SALIENCE. The examples we use are adverbial attachments and relative clause attachments, including the interesting cross-linguistic data we encountered in Section 2.2.2. We begin this section with adverbial attachments.

Recall that in English, *MOD >> SALIENCE >> *H-D. Note that SALIENCE does not distinguish between competing adverbial attachments, for it only marks attachments to NP's. Furthermore, none of the adverb attachment ambiguities we found in the literature involved candidates taking more than one adjunct. Thus, in each of the examples we provide, *H-D will be responsible for the attachment preferences that arise. We run through a small number of examples for illustrative purposes. However, in each case, the most optimal parse will be the one exhibiting local attachment.

Consider first the following sentence taken from Vervaeke (2000):

(58) John said the car crashed last night.

This sentence is ambiguous, for the AdvP 'last night' may attach to either of the preceding VP's.[34] The structure that attaches 'last night' to the VP headed by 'crashed' incurs no violations of *H-D, for there are no intervening maximal projections between the VP and the AdvP. The structure that attaches the AdvP to the VP headed by 'said' incurs four violations of *H-D, for there are four maximal projections intervening between the VP headed by 'said' and the AdvP (CP, IP, NP, VP). Thus, the low attachment is predicted to be preferred, which is found to be the case (Vervaeke, 2000).

Let us examine now whether our system can handle the cross-linguistic data that were presented in Section 2.2.2. Recall sentence (24), which we rewrite here as (59):

(59) Bill said John died yesterday.

There is an ambiguity at the word 'yesterday', which can attach to either of the preceding VP's. Again, if the AdvP attaches to the VP headed by 'died', it incurs no violations of *H-D. If it attaches to the VP headed by 'said', it incurs four violations of *H-D (CP, IP, NP, VP). Thus, the system predicts that the low attachment is preferred over the high attachment, and this is indeed observed to be the case (Gibson et al., 1996).

The Spanish translation of (59) is given below:

(60) Juan dijo que Bill se murió ayer.

Again, *H-D is left to account for the ambiguity resolution and, just as in the English version above, it prefers low attachment, for there are less maximal projections intervening between the local VP and the AdvP than between the non-local VP and the

[34] Note that this is a Type-2 ambiguity. However, (it is being assumed that) PROB-ACC does not favour either of the two possibilities, and thus it is up to the other constraints to determine the optimal parse. Each of the following examples is a Type-2 ambiguity that PROB-ACC does not resolve. Thus, we will no longer mention the 'Type' of the ambiguity.

AdvP. Hence, the system predicts that in both English and Spanish (indeed in all languages) there should be an observed preference to attach adverbs locally.[35]

What about the relative clause attachments that were discussed in Section 2.2.2? We saw that relative clause attachment preferences undergo a curious transformation when moving from two-NP site ambiguities to three-NP site ambiguities. In the structure NP_1 NP_2 RC, there is a preference for the RC to attach to the local NP in English and to the non-local NP in Spanish. In the structure NP_1 NP_2 NP_3 RC, the order of attachment preference in *both* English and Spanish is $NP_3 >> NP_1 >> NP_2$. Is our system able to capture this phenomenon?

Before we analyse the ability of our system to account for the above preferences, let us recall the process that led to the development of our constraint system. Our goal was the development of an optimality theoretic system of constraints that would be able to represent the behaviour of the HSPM. We took our start from the OT systems of Gibson and Broihier (1998), arguing that they were unsuccessful primarily because they did not incorporate well established psycholinguistic results (such as lexical frequency information). We then dove into the psychological literature to determine what factors or constraints are involved in the workings of the HSPM. It was determined that lexical frequency information, working memory limitations, and salience of discourse entities (for guiding anaphoric attachments) are the factors that guide parsing preferences. We translated these factors into the following set of OT constraints, {PROB-ACC, SALIENCE, *MOD, *H-D}, providing associated markedness calculi for each constraint. It was further argued, again based on the psychological finding that lexical frequency information is the most important constraint in parsing decisions, that PROB-ACC outranks each of *MOD, SALIENCE and *H-D. Typology was argued to be a result of permutations of the ordering of the constraints SALIENCE , *MOD and *H-D. This constraint system is argued to be a faithful representation of the psychological data. It captures, in essence, that and only that which has presented itself through empirical testing and rational analysis as being involved in parsing. The goal throughout has been

[35] Assuming, of course, that PROB-ACC and *MOD are neutral about attachment preferences.

to show that Gibson and Broihier (1998)'s conclusion, viz. that 'standard OT' is unable to account for the HSPM, is incorrect. We have tried to produce a system of constraints obeying the property of strict domination that is able to account for all the data. It has, thus far, captured an impressive array of results. But the question remains: Is it enough?

Recall sentence (27) (cf. Section 2.2.2), rewritten below as (61):

(61) The journalist interviewed the daughter of the colonel who had had the accident.

The point of ambiguity occurs at the word 'who', for it can modify either 'the daughter' or 'the colonel'. Let X be the former parse, and let Y be the latter parse. Below, we present the constraint violations incurred by each, with the analysis of X in (62) and the analysis of Y in (63).

(62) *MOD Violations: 1 (because 'the daughter' has two modifiers, 'of the colonel', and 'who had had the accident')
SALIENCE Violations: 2 (one for 'recency', as one NP ('the colonel') is passed over in making the attachment; one for grammatical obliqueness, as 'the daughter' is a direct object)
*H-D Violations: 4 (one each for the intervening maximal projections PP, NP ('the colonel'), CP, NP (the gap))

(63) *MOD Violations: 0 (because each NP has one adjunct)
SALIENCE Violations: 3 (all from grammatical obliqueness, because local NP belongs to an adjunct specifying 'which daughter')
*H-D Violations: 2 (one for the intervening CP, and one for the intervening NP (the empty element))

Recall that in English, the constraint ranking is *MOD >> SALIENCE >> *H-D. The structure exhibiting local attachment (63) incurs less violations of *MOD than the structure exhibiting non-local attachment (62). Hence, our system predicts that in two-

NP relative clause ambiguities, local attachment is preferred. This indeed matches the empirical results discussed in Section 2.2.2.

The Spanish translation of (61) is given below in (64):

(64) El periodista entrevisto a la hija del coronel que tuvo el accidente.

The observed preference in Spanish is to attach high, so that it is 'la hija' who had the accident, not 'del coronel'. Let X be the structure where the relative clause attaches high, and let Y be the structure where the relative clause attaches low. We present analyses of the violations that X and Y incur in (65) and (66), respectively.[36]

(65) *MOD Violations: 1 (because 'la hija' has two modifiers)
SALIENCE Violations: 2 (one for recency, as one NP ('del coronel') is passed over in making the RC attachment; one for grammatical obliqueness, as 'la hija' is a direct object)
*H-D Violations: 4 (cf. Footnote 34)

(66) *MOD Violations: 0 (because each NP has one modifying adjunct)
SALIENCE Violations: 3 (all from grammatical obliqueness, as attachment is to an adjunct)
*H-D Violations: 2 (cf. Footnote 34)

In Spanish, the constraint ranking is SALIENCE >> *MOD >> *H-D. As the structure exhibiting high attachment (65) incurs less violations of SALIENCE than the low attachment structure (66), our system predicts that in the structure NP_1 NP_2 RC, native

[36] It is useful to note that in the sequence NP_1 NP_2 RC, attachment to the more local NP results in two violations of *H-D, and attachment to the non-local NP results in four such violations. See the English example above for specifics. In fact, for any sequence NP_1 NP_2 … NP_k RC, attachment of the RC to NP_i results in 2*[k-(i-1)] violations of *H-D. This has the effect of adding two violations for each NP that is passed over in making an attachment. The two added violations correspond to the extra NP and PP that intervene between the maximal projection headed by NP_i and the RC.

Spanish speakers should prefer attachment to NP_1 rather than to NP_2. This is indeed the case (cf. Section 2.2.2, Gibson et al. 1996, Hemforth et al. 2000).

We now move on to the crucial three-NP relative clause attachment ambiguities. The theory should be able to predict that in both English and Spanish, the order of attachment preference is $NP_3 >> NP_1 >> NP_2$. For preciseness, we analyse sentence variants of (29), rewritten below as (67) and (68).[37] However, the results are not specific to these examples, for the constraints are structural, and hence independent of the actual content of the lexical items.

(67) Pedro turned on [NP_1 the lamp near [NP_2 the painting of [NP_3 the house]]] [CP that was damaged in the flood].

(68) Pedro encendió [NP_1 la lámpara cerca de [NP_2 la pintura de [NP_3 la casa]]][CP que fue dañada en la inundación].

The reader may have noticed that the distribution of constraint violations is the same in both English and Spanish (cf. The remark above about the content-independent nature of the constraints). Hence, to simplify our analysis, we consider the abstract structure NP_1 NP_2 NP_3 RC, and describe the constraint violations attributed to the structures where the RC attaches to NP_1, to NP_2, and to NP_3. We do this in (69) to (71).

(69) Attachment to NP_1: 6 Violations of *H-D (cf. Footnote 34)
3 Violations of SALIENCE (one violation for grammatical obliqueness, as NP_1 is a direct object, and two violations for recency preference, as two NP's (NP_2, NP_3) are passed over in making the attachment)
1 Violation of *MOD (as NP_1 has two modifiers, viz., NP_2 and

[37] By 'variants' I mean that what is measured is the 'goodness' not of the sentences as they appear here, but rather of disambiguated versions of the above sentences, where only one NP site is available for attachment (by number agreement in this case). The 'goodness' of the disambiguated versions is measured by on-line grammaticality judgements and reading times.

RC)

(70) Attachment to NP_2: 4 violations of *H-D (cf. Footnote 34)
4 violations of SALIENCE (three violations for grammatical obliqueness, as NP_2 is an adjunct, and one violation for recency preference, as one NP (NP_3) is passed over in making the attachment)
1 Violation of *MOD (as NP_2 has two modifying adjuncts, viz., NP_3 and RC)

(71) Attachment to NP_3: 2 violations of *H-D
3 violations of SALIENCE (three violations for grammatical obliqueness, as NP_3 is an adjunct)
0 Violations of *MOD (as each NP has only one modifier)

We begin our analysis with English. Recall that the constraint ranking in English is *MOD >> SALIENCE >> *H-D, and that we want to predict (in both English and Spanish) that (71) >> (69) >> (70). Structure (71) fares the best in terms of *MOD violations, so we have correctly predicted the most preferred structure. However, note that (69) and (70) are tied with respect to *MOD. However, (69) fares better than (70) with respect to SALIENCE, and hence our system correctly predicts that (71) >> (69) >> (70).

In Spanish, the constraint ranking is SALIENCE >> *MOD >> *H-D. In this case, (69) and (71) are tied with respect to SALIENCE, each incurring three violations. However, (71) fares better than (69) with respect to *MOD, the next highest constraint. Hence, (71) >> (69). Furthermore, as both (69) and (71) incur less violations of the high ranked SALIENCE than (70), we get the predicted ordering: (71) >> (69) >> (70).

Our constraint system is therefore able to capture the observed relative clause attachment preferences in both the two-NP site ambiguities and the three-NP site ambiguities. More generally, we have seen that our system captures a large set of parsing preferences across a broad range of ambiguities. However, a theory of the HSPM should not only be able to account for the processing of ambiguous sentences, but should also describe the processing complexity associated with the processing of unambiguous sentences. We briefly turn our attention now to such structures.

4.2.3 Centre-Embedding

We are essentially getting this part of our work for free. Gibson and Pearlmutter (1998) propose head-dependent distance as an aspect of locality that can account for the processing complexity of unambiguous sentences. Our work above is an attempt to show that head-dependent distance is all there is to locality; we need no other notion of locality above and beyond head-dependent distance in order to account for the observed data. We now present some simple examples to show how the constraint *H-D is able to explain the fact that certain unambiguous syntactic structures, namely centre-embedded structures, are more difficult to process than others. The reader is referred to Gibson and Pearlmutter (1998) for further discussion of head-dependent distance and the complexity of nested structures.

Consider the following two sentences, taken from Lewis (2000):

(72) The bird chased the mouse that scared the cat that saw the dog that ate the pumpkin that grew in the garden.

(73) The salmon that the man that the dog chased smoked fell.

Sentence (72) contains four embedded clauses. These are right-branching structures and are relatively easy to comprehend. Sentence (73) contains two centre-embedded structures and, to most readers, is nearly incomprehensible. What is the reason for this complexity? The *H-D constraint explains this quite easily.

In (72), the head noun and its dependent gap are always two maximal projections apart.[38] For example, in the clause 'the mouse that scared the cat', 'the mouse$_i$' and the gap 'e_i' (that scared the cat) are separated by maximal projections CP and IP. This does not cause a great burden on working memory, for only two violations of *H-D are incurred.

In (73), 'the man$_j$' and the gap 'e_j' (where 'the dog chased $e_{j'}$) have four intervening maximal projections: CP, IP, NP and VP.[39] Similarly, 'the salmon$_i$' and the co-indexed empty element 'e_i' are separated by nine intervening maximal projections, CP, IP, NP, CP, IP, NP, VP, NP and VP. This poses far too great a burden on working memory, indicated by the large number of *H-D violations incurred by this structure.

Observe that the framework we provide has the potential to make very precise predictions about processing breakdown. For example, with enough data analysis, it should be possible to make a generalisation of the form: If a structure X incurs m or more violations of *H-D, the HSPM will suffer from processing breakdown. We have seen above that two violations are manageable, whereas nine are not. By analysing enough data, an upper bound should be deducible declaring the number of *H-D violations that the HSPM is able to accommodate. Furthermore, note that because these sentences are unambiguous, SALIENCE does not operate here, for anaphoric attachments have only one possible resolution. Additionally, none of the above structures incurs any violations of *MOD, for each XP is modified by at most one adjunct. Thus, the model accounts for the complexity of processing centre-embedded structures by appealing to the constraint *H-D.

[38] We parse this sentence as follows: [IP[NP the bird][VP[V chased][NP[NP$_i$ the mouse][CP[C that][IP[NP e_i][VP[V scared][NP[NP$_j$ the cat][CP[C that][IP[NP e_j][VP[V saw][NP[NP$_u$ the dog][CP[C that][IP[NP e_u][VP[V ate][NP[NP$_w$ the pumpkin][CP[C that][IP[NP e_w][VP[V grew][PP in the garden]]]]]]]]]]]]]]]]]]]].

5. Discussion

In this final chapter, we provide some concluding remarks on a wide array of topics. They are all comments on the work pursued in this thesis, with some looking back at older issues to note how our system relates to those issues, and with others looking to future issues that arise as a result of our work. Before involving ourselves in discussion of such matters, I would like to take this opportunity to list, in bullet form notes, what I take to be the main achievements of our work:

- The development of a sophisticated theory of the HSPM that improves upon the state of the art. It does so by providing a very honest theoretical representation of empirical results.
- It is distinguished from most current theories of the HSPM in that it precisely characterises (aspects of) the cognitive architecture that work with the HSPM in producing fast, reliable parses of linguistic inputs. For example, it isn't satisfied with just stipulating that lexical frequency information influences parsing: it tells us how that information is represented, and how it influences the HSPM (through PROB-ACC).
- It contributes to optimality theory by providing the most comprehensive analysis of parsing within the OT literature to date. Hence, it continues the expansion of OT into the domain of language processing. It also justifies our intuitions that OT should provide a suitable framework within which to formulate theories of the HSPM, therefore refuting GB's conclusion.
- It provides functional explanations for each constraint, illustrating how each constraint is a necessary contributor to the computational efficiency of the HSPM. Thus, the theory is not only descriptive, but explanatory also.
- The work unifies a range of ambiguities into a small, compact set of ambiguity types, thereby adding efficiency to the HSPM's ambiguity resolution methods.
- It helps situate the language faculty within the broader cognitive system, an element which has been lacking in the OT work to date.

---

[39] We parse this sentence as follows: [IP[NP[NP_i the salmon][CP[C that][IP[NP[NP_j the man][CP[C that][IP[NP the dog][VP[V chased][NP e_j]]]]][VP[V smoked][NP e_i]]]]][VP fell]].

- The theory makes very clear, falsifiable predictions, and is therefore well constrained.

5.1 Adequacy of the Theory

The true measure of any theory is its ability to describe and predict data. In Chapter 4 we had put our theory to the test and demonstrated that it is able to capture a large number of experimentally determined parsing preferences. Of course, this is by no means an exhaustive set of test cases, but the range and number of ambiguities covered is quite impressive. The theory was found to predict, in addition to preferences in ambiguity resolution, the processing complexity of unambiguous centre-embedded structures, hence rendering the HSPM not just as a theory of ambiguity resolution, but as a general theory of sentence processing, as it should be.

The work advanced in this thesis raises some important questions that should provide an impetus for future work. We have reduced various structural ambiguities to Type-1, Type-2 and Type-3 ambiguities. Surely, this cannot be an exhaustive reduction. By enumerating the full set of structural ambiguities found in natural language, and by attempting to find the common structure underlying them all, there is a good chance that we can have a streamlined set of ambiguity types that the HSPM deals with. By integrating the seemingly unconnected ambiguities found in natural language, we can develop theoretically elegant parsing systems that a) can account for the efficiency of the human sentence processor and b) can be fruitfully implemented in various computational domains, such as ambiguity resolution, speech recognition, etc.

A serious flaw in our work, however, is that we have only accounted for a subset of *English* ambiguities. Surely, this covers only a fraction of the ambiguity types that abound in the world's languages. Only by pursuing more cross-linguistic work will we be able to truly make significant progress. It's like studying genetics by restricting examination to one species; only the shallowest of knowledge can be gained by such methods of inquiry. The psycholinguistic community is perhaps distinguished within the broader linguistics community in that it has predominantly focussed on English. This

will continue to hold back progress until serious cross-linguistic work is carried out, as it has in other domains, such as syntax.

5.2 Preferences, Not Rules

The constraints that we have outlined above are not hard constraints that force interpretations one way or another. Rather, the constraints give rise to preferences that are more gradable than, say, grammaticality judgements. For example, when we say that in two-NP relative clause attachment ambiguities Spanish speakers prefer to attach high, this does not mean that they must attach high, nor does it mean that low attachments are exceedingly difficult. It means that, more often than not, most Spanish speakers will interpret such a sentence as a high attachment, and that more often than not most Spanish speakers' reading times will be quicker for high attachments. Thus, the constraints guiding the HSPM are of a different kind than those guiding universal grammar (hf. UG). Hence, we maintain a competence/performance distinction in our work. However, that is not to say that there are not interesting relations between competence systems and performance systems. We discuss such relations in Section 5.7.

5.3 Functional Explanation

Each one of the constraints in our system has been introduced with a specific purpose in mind: to make the HSPM computationally efficient. PROB-ACC makes the HSPM a safe-betting goody-two-shoe, but a fast and reliable one. When faced with local ambiguities, PROB-ACC guides the HSPM to the most probable interpretation of the input. This occurs quickly, and results in the HSPM being right most of the time. *MOD and *H-D serve to minimise the strain on the computational resources of the HSPM by yielding a preference to work locally. SALIENCE helps the HSPM to make 'good' anaphoric attachments by ordering the set of possible attachments according to their 'goodness', or salience. All the constraints thus help the HSPM in its search for an optimal parse by biasing it toward certain structures. Connecting all these ideas together is the underlying assumption that the HSPM is incremental, using immediate analysis in parsing the input it receives. These ideas provide an answer to the question, what makes the HSPM so computationally efficient?

5.4 Reanalysis

In our system, reanalysis occurs by backtracking to the previous point P in the parse that exhibits a structural ambiguity, and by doing another evaluation at P of the candidate set minus the structure previously selected optimal. That is, if {A, B, C,…} is the set of candidates ordered by harmony, and A was previously selected optimal, then the new candidate set will be {B, C,…} and B will be chosen. This notion of reanalysis derives from the assumption that at each point in the parse, GEN outputs only those candidates that are grammatical extensions of previous parses. In particular, the current candidate set cannot include reanalyses of previous parses.

Alternatively (Smolensky, 1997), one may adopt a system whereby at each point in the parse, GEN outputs all grammatical parses of the given input, including reanalyses of previous parses. To maintain the result that optimal structures are (usually) extensions of previous parses, one simply introduces faithfulness constraints marking against structures that reanalyse the previous parse.

The system as outlined by Smolensky (1997) is more attractive than ours in the sense that it is far more general. At all points in the parse, it allows GEN to output all grammatical interpretations of the given input, whereas our system imposes the (somewhat artificial) restriction that only grammatical extensions of the previous parse may be output. However, the incrementality of the HSPM is more transparent in our formulation than in Smolensky (1997), which adds a measure of justification to the 'extensionality' restriction. Ultimately, it would be attractive to adopt the generality of Smolensky (1997) while maintaining 'extensionality' in a far stronger way than just marking against it with the use of a faithfulness constraint. How to do this remains a prospect for future work.

5.5 Production

Are the constraints involved in sentence comprehension the same as those involved in production? Our text has been filled with the acronym 'HSPM', thereby implicating the constraint system as an interpretive system that is focussed on describing how readers

and listeners process linguistic inputs. Assuming Gricean maxims of reasoning, speakers/writers should produce linguistic outputs that maximise their chances of being understood, and hence that maximally satisfy the constraints of the HSPM. However, it is difficult to see how a constraint like 'SALIENCE' may guide production. The essence of that constraint is to help readers/listeners resolve ambiguities by making 'good' attachments. To think that a speaker is producing her utterances attempting to maximally satisfy the SALIENCE constraint would mean that, in many cases, she would have had to already mapped the form her expression will take before having uttered it. This strikes us as being highly unlikely, for it doesn't account for such things as the spontaneity of speech and, perhaps more severely, contradicts the incrementality of the HSPM. Although we take the HSPM to be predictive, it only follows a 'look-ahead' of one word, not an entire set of words. Hence, we reserve the constraint set outlined above to be solely descriptive of the sentence comprehension mechanism. However, this does not exclude the possibility of interesting relations existing between production systems and comprehension systems. We are only arguing against equating the two.

5.6 Language and Cognition

The work put forth in this thesis has attempted to situate the language faculty within a general cognitive setting. This does not mean that we do not consider the language faculty to be modular, for we do. Rather, it means that what we know about human cognition should inform our theories of language, and our theories of language should inform our theories of cognition. For example, a theory of language processing that contradicts what we know about working memory will be wrong before it even has a chance to get off the ground. In effect, this is a methodological argument for the existence of cognitive science, whereby scientific and philosophical inquiries from several domains work to mutually constrain and inform each other. Our theory of the HSPM, for example, has benefit enormously by taking into account very simple psychological results, such as the incrementality of the HSPM, the restrictions imposed by working memory, and the fact that subcategorization frequencies influence parsing.

5.7 Grammars and Parsers

The assumption that there is an innate system of grammatical constraints or principles is quite uncontroversial in linguistic theory. Indeed, much of modern linguistic research is driven by this very assumption. In contrast, the assumption that there is a universal parser is not so well established in psycholinguistics.

Our view on the matter is that there is a universal parser, viz., the set of constraints {PROB-ACC, *MOD, *H-D, SALIENCE}. This set characterises the parser in all languages, modulo ranking permutations. That there should be such a universal system is supported by several considerations. First, one of the most striking aspects of human language is that all (healthy) humans in all languages process sentences quickly and reliably. If there were ever a candidate for a 'linguistic universal', few come to my mind that are more obvious than this. This universality suggests that all humans share some mechanism that they use to cut up the sound sequences they hear into well-formed syntactic structures. Second, the computational efficiency of language processing arises out of requirements on interpretation that themselves are based on the restrictions imposed by general cognitive architecture, such as working memory limitations. As such cognitive conditions are (presumably) universal, the parsing mechanism that is derived from them should similarly be universally invariant. Third, to actually acquire a language in the first place, one needs to be able to process linguistic inputs to determine, for example, their syntactic structures, the roles of various arguments (subject, object), etc. It is difficult to see how language qua grammar could be acquired without some innate mechanism that parses the input language in the first place. This suggests that some mechanism is already (innately) in place working in tandem with the grammatical component of the language acquisition device to help the language faculty develop. These arguments taken singularly do not prove our claim that there is a universal parser, but the conjunction of each gives a lot of support to the claim. But if there is such a universal mechanism, and if our set of constraints describes the mechanism correctly, then we must answer the question, why is there cross-linguistic variation in the constraint rankings of HSPM?

We assume that there is a strong relation between the grammar and the parser, an assumption that gives us a great deal of explanatory power. We claim that the constraints in the grammar give rise to certain syntactic structures in the linguistic environment, and that the HSPM ranks its constraints in order to maximise processing efficiency of the structures in the environment. In our case, recall that PROB-ACC is universally the highest ranked constraint in the constraint set. In Spanish, we saw that SALIENCE >> *MOD >> *H-D, whereas in English *MOD >> SALIENCE >> *H-D. What is the explanation behind this ranking?

Hemforth et al. (2000) point out that English makes inconsistent use of relative pronouns, often omitting them or replacing them with the use of a generalised complementizer. For example, both of the following are acceptable English sentences:

(74) The daughter of the teacher that Peter visited…
(75) The daughter of the teacher Peter visited…

The reasoning goes that because of its inconsistent use of relative pronouns, English cannot rely on anaphoric processes for relative clause attachments. As a result, the need for a strong SALIENCE constraint becomes reduced. A language that is stricter about its use of relative pronouns, like Spanish, relies on SALIENCE for purposes of anaphoric attachment to a greater degree, hence increasing the need for a strong ranking of the SALIENCE constraint.

Using the same sort of reasoning as above, as well as that found in Gibson et al. (1996), we would predict that languages with SVO and OVS word order should have a weak *H-D constraint. This is because the distance between verbs and their arguments is already taken care of by word order constraints. Languages with OSV, SOV, VSO or VOS word order should be predicted to have a higher ranked *H-D constraint. In both English and Spanish, the canonical word order is SVO, although Spanish is freer with regards to variability in maintaining that order as it also allows a VOS order. In each language, we have found that *H-D is low ranked, thus matching our predictions. We suggest that by

analysing word order and strictness of pronoun usage in different languages, we should be able to motivate constraint rankings of the HSPM in those languages.

Going in the other direction, by positing a strong relation between the grammar and the parser, we see how functional motivations can give rise to constraints in the grammar. For example, the *H-D constraint marks against head-dependent distance. This can be used as an explanation for why we don't see verbs in languages that take, say, twelve arguments. Why shouldn't there be such verbs in a language? If there can be verbs with one argument, with two arguments, with three arguments, why not verbs with twelve arguments? The question may sound slightly silly, but is worth asking. The answer that comes out of our work is that such a structure would result in heavy computational complexity for the HSPM resulting in processing breakdown. Thus, the existence of *H-D in the HSPM prevents the existence of verbs with a large number of arguments. Without assuming this strong relationship between the grammar and the parser, such explanations become an almost vanishing possibility. Many interesting explanations and predictions such as these are readily forthcoming by assuming a close grammar-parser connection.

5.8 What Must a Theory of Sentence Processing Do for Us?

Crocker (1996, p. 33-34) states several basic requirements that any theory of the HSPM must satisfy. In this section we list these requirements in question form and indicate how the theory developed in this work answers the given questions.

Question 1: How are the rules of grammar used to construct an analysis of an utterance?
Answer 1: At each step in the parse, GEN outputs all grammatical extensions of the previous parse, and selects the candidate that is optimal over the metric defined by the HSPM, viz., the constraint set {PROB-ACC, SALIENCE, *MOD, *H-D}.

Question 2 (accounting for empirical results):

Question 2a: How do humans parse sentences so rapidly?

Answer 2a: Humans are endowed with an innate, universal set of constraints that are designed to make the parsing process computationally fast. The constraints can be thought of as providing instructions to the HSPM to select a particular candidate:

- PROB-ACC: Select the most likely parse, i.e. select that candidate whose structure is consistent with the highest ranked element in the relevant lexical order.
- SALIENCE: Make anaphoric attachment to that item highest ranked in the set of discourse entities ordered by salience.
- *H-D: Select that structure that minimises the distance between heads and dependents.
- *MOD: Select that structure that minimises the number of 'excessive' adjuncts.

Each of these constraints speeds up the HSPM's computations by reducing the load on working memory and by biasing the parse toward a particular structures. Thus, the HSPM is employed with heuristic search strategies that avoid excessive combinatorial operations in the search for an optimal parse.

Question 2b: What leads to preferred readings in ambiguous sentences?

Answer 2b: Those candidate structures that are more harmonic with respect to the constraint set are more preferred.

Question 2c: Why are some sentences more difficult to process that others?

Answer 2c: Because they are less harmonic with respect to the constraint set defining the HSPM.

Question 2d: How do humans recover from errors made during parsing?

Answer 2d: By backtracking to the previous point of ambiguity, and selecting the optimal structure out of the candidate set at that point *minus* the structure that was followed that led to the parsing error. In other words, on the previous pass, if {A, B, C,…} was the (ordered by harmony) candidate set, then this time B is chosen.

Question 2e: What causes processor breakdown, or 'garden path' phenomena?

Answer 2e: The HSPM is incremental. At a particular point in the parse, the candidate structure that is optimal over the HSPM constraint metric is not the correct structure. This follows immediately from the immediacy principle. It is only later on down the parse that the HSPM realises an error was made. This results in the HSPM backtracking to fix the error, as discussed in Answer 2d.

Question 2f: What leads to over-loading in non-ambiguous structures?
Answer 2f: Over-loading results when the number of violations of the HSPM constraints passes a certain threshold. We haven't yet identified what that threshold is, but we saw that in the processing of centre-embedded clauses, the processing complexity was directly related to the number of *H-D violations incurred. We conjecture that over-loading in non-ambiguous structures is most likely particularised to high numbers of violations of the constraints that are directly derived from working memory considerations, viz., *MOD and *H-D. For example, let X and Y be two non-ambiguous structures that are otherwise structurally similar, and whose total number of constraint violations is z. Suppose X distributes the z violations evenly amongst all four of the HSPM constraints, while Y divides them between *MOD and *H-D only. We conjecture that structure Y should lead to greater processing complexity than structure X.

5.9 A Look to the Future

We use this section to indicate some possible directions for future research.

First, as we outlined above (cf. Section 5.1), a continued classification of ambiguities into a small set of ambiguity types would yield nice generalisations about the range of ambiguities present in natural language, and the aspects of the HSPM that allow it to handle such ambiguities. Of course, to obtain full generality, cross-linguistic work is sorely needed.

Second, we saw that lexical frequency information is highly important in determining parsing preferences. A natural question to ask is, how, if at all, does lexical frequency information change over time? To what extent is this change dependent on

environmental factors, and to what extent might this change be just a simple matter of cognitive decay? For example, suppose that for a speaker A, the word 'claim' is more likely to take a sentence complement, as in 'I claimed the victory was mine' rather than an NP argument, as in 'I claimed the victory'. Now, suppose that A were inserted into an environment where ninety-five percent of the time 'claim' was followed by a simple NP. I conjecture that even if the total number of times speaker A had come across 'claim(NP)' would be less than 'claim(S)', at some point, when put under the same experimental situations as those discussed in this work, speaker A would begin to prefer 'claim(NP)' over 'claim(S)'. If this is true, then it isn't just frequency information alone that is relevant here. It is probably more accurate to claim that ordered subcategorization information is represented by a two-dimensional function of both 'recency of exposure' and overall frequency.

Third, how is this lexical frequency information learned? Is it a product of the constraint ranking? Is there some separate learning mechanism dissociated from the language acquisition device that acquires this information? How are we to determine this? A fundamental principle of optimality theory offers us a clue. The principle of 'richness of the base' says that:

> *the set of all inputs to the grammars of all languages is the same…Richness of the base requires that systematic differences in inventories arise from different constraint rankings, not different inputs. The lexicon of a language is a sample from its inventory:* ***all systematic properties of the lexicon thus arise indirectly from the grammar…***
>
> (Smolensky 1996, p.3, emphasis added)

By interpreting 'systematic' as meaning 'robust' and 'revealed reliably under experimental tasks', we may take the lexical probability information as being a systematic property of the lexicon. This would mean that there should be something in the grammar that would yield the probability values. This raises the following question: What kind of relation exists between the grammar and the lexicon that would allow

probability information associated with lexical entries to arise out of the grammatical system of constraints? No obvious answer is forthcoming.

Alternatively, perhaps it is a mistake to interpret the lexical frequency information as being a 'systematic' property of the lexicon. Unfortunately, we have no way of determining whether it is a mistake to make such an interpretation because the notion 'systematic property of the lexicon' has been left vague and unclear in the OT literature. As such, it is difficult to determine the best way to proceed to give an answer to the question, how is the lexical frequency information learned?

5.10 OT and the HSPM

What has our study of optimality theory and the human sentence processing mechanism taught us? It has revealed that, as the structural similarities suggest, OT is well suited to accounting for the behaviour of the HSPM. By simply accommodating psycholinguistic results into our theoretical formalism, we have developed a descriptively powerful, well-constrained, elegant theory of the HSPM. It improves upon current theories along significant dimensions discussed in the text, and opens many new doors for future research. The OT-HSPM merger proposed here benefits OT because it expands optimality theoretic empirical coverage to new domains. It benefits theories of the HSPM by allowing them to take advantage of the simplicity of the optimality theoretic framework. We gain many powerful theoretical tools by working within OT, such as an elegant markedness calculus and a symmetric notion of typology (simple permutations). It is hoped that the ideas developed here give rise to further work unifying OT and language processing.

References

Baddeley, A.D. (1986). *Working Memory*. Oxford: Oxford University Press.

Blutner, R. (2001). *Optimality Theory*. Lecture notes on optimality theory delivered at the Dutch Research School in Logic, October, Nunspeet, Holland.

Boland, J.E. (1997). Resolving Syntactic Category Ambiguities in Discourse Context: Probabilistic and Discourse Constraints. *Journal of Memory and Language* 36: 588-615.

Carpenter, P.A., Miyake, A. and Just, M.A. (1995). Language Comprehension: Sentence and Discourse Processing. *Annual Review of Psychology* 46: 91-120.

Carroll, D.W. (1999). *Psychology of Language, Third Edition*. Toronto: Brooks/Cole Publishing Company.

Crocker, M. (1996). *Computational Psycholinguistics*. Dordrecht: Kluwer Academic Publishers.

Cuetos, F. and Mitchell, D.C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in Spanish. *Cognition* 30: 73-105.

Fanselow, G., Schlesewsky, M., Cavar, D. and Kliegl, R. (1999). Optimal Parsing, Manuscript, University of Potsdam.

Frazier, L. and Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14: 178-210.

Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E. and Hickok, G. (1996). Recency preference in the human sentence processing mechanism. *Cognition* 59: 23-59.

Gibson, E. and Broihier, K. (1998). Optimality Theory and the Human Sentence Processing Mechanism. In Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky (eds.), *Is the Best Good Enough? Optimality and Competition in Syntax,* Cambridge, MA: MIT Press, 157-191.

Gibson, E. and Pearlmutter, N. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences* 2: 262-268.

Hemforth, B., Konieczny, L. and Scheepers, C. (2000). Syntactic Attachment and Anaphor Resolution: The Two Sides of Relative Clause Attachment. In Matthew Crocker, Martin Pickering, and Charles Clifton Jr. (eds.), *Architectures and Mechanisms for Language Processing,* Cambridge: Cambridge University Press,

259-281.

Lewis, R. (2000). Specifying Architectures for Language Processing: Process, Control, and Memory in Parsing and Interpretation. In Matthew Crocker, Martin Pickering, and Charles Clifton Jr. (eds.), *Architectures and Mechanisms for Language Processing*, Cambridge: Cambridge University Press, 56-89.

McCarthy, J. (2002). *A Thematic Guide to Optimality Theory*. Cambridge: Cambridge University Press.

Prince, A. and Smolensky, P. (1997). Optimality: From Neural Networks to Universal Grammar. *Science* 275, 1604-1610.

Rayner, K., Carlson, M. and Frazier, L. (1983). The Interaction of Syntax and Semantics During Sentence Processing: Eye Movements in the Analysis of Semantically Biased Sentences. *Journal of Verbal Learning and Verbal Behavior* 22: 358-374.

Shapiro, L.P., Nagel, H.N. and Levine, B.A. (1993). Preferences for a Verb's Complements and Their Use in Sentence Processing. *Journal of Memory and Language* 32: 96-114.

Smolensky, P. (1996). The Initial State and 'Richness of the Base' in Optimality Theory. Technical Report No. JHU-CogSci-96-4. Baltimore: Department of Cognitive Science, Johns Hopkins University.

Smolensky, P. (1997). Optimal Sentence Processing, Manuscript, Johns Hopkins University.

Taraban, R. and McClelland, J. (1988). Constituent Attachment and Thematic Role Assignment in Sentence Processing: Influences of Content-Based Expectations. *Journal of Memory and Language* 27: 597-632.

Trueswell, J.C., Tanenhaus, M.K. and Kello, C. (1993). Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preferences from Garden Paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19: 528-553.

Trueswell, J.C., Tanenhaus, M.K. and Garnsey, S.M. (1994). Semantic influences on parsing: use of thematic role information in syntactic disambiguation. *Journal of Memory and Language* 33, 285-318.

Trueswell, J.C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language* 35, 566-585.

Vervaeke, J. (2000). Lecture, 13 June, University of Toronto, Toronto, Ontario, Canada.