University of Utrecht

BA Thesis Linguistics

How Parsing Operator Chunking predicts Reading Times in Sentence Processing

Evidence from the Natural Stories Corpus for Hale's application of the Chunking Theory of Learning on Sentence Parsing

Student:
Elze van der Werf
e.g.vanderwerf@students.uu.nl
Student ID: 6203116

Supervisor:
Dr. Jakub Dotlačil
j.dotlacil@uu.nl

Second evaluator:
Dr. Rick Nouwen
R.W.F.Nouwen@uu.nl

Bachelor of Linguistics
Department of Languages, Literature and Communication
Faculty of Humanities
26 Jun, 2020

Abstract

Elaborating on earlier theories in the field of sentence processing and parsing strategies, Hale (2014) proposed that the Chunking Theory of Learning (CTL) might be considered a good potential for relating a concrete mechanism to the sentence complexity metric Surprisal (Hale, 2001), which provides a mathematical specification of the probability of the next word in the sentence. Applying CTL to sentence processing, Hale assumed that parsing operators can be fused together into a quicker executing macro-operator if used more often, resulting in faster parsing for more familiar sentence structures, at the same time reducing surprisal effects. The present study provides an examination of Hale's theory on parsing action chunking, testing its predictions on the Natural Stories Corpus (Futrell et al., 2018). The results show a correlation between cohesion degree of parsing operator trigrams and average reading times, supporting the idea that parsing action tuples can be learned to be a chunk. We will conclude that the presented results are in line with Hale's predictions, and that further research should give insight into possible internal or external effects at play.

Keywords: sentence processing, parsing, surprisal, chunking, self-paced reading times

Contents

Introduction

How are human beings so proficient in understanding spoken and written sentences effortlessly? In this comprehension process many informational features are involved, ranging from phonetic and morphological features, sociolinguistic aspects and semantic interpretation to syntactic structure.

This study's focus lays on the grammatical knowledge involved in sentence processing. Psycholinguists continue being puzzled about humans' overwhelming skill to build syntactic representations of sentences. Experiments on sentence processing difficulty have led to theories such as the Sausage Machine theory by Frazier and Fodor (1978), various formal parsing algorithms such as Generalized Left-Corner parsing (Demers, 1977) and to the well-known Garden-Path Theory (Frazier & Rayner, 1982), which gives a heuristic account for processing difficulties in parsing garden-path sentences. More recently, a different perspective on syntactic processing has become popular, finding its roots in the information theory in the tradition of Shannon (1948), in which sentence processing difficulties are related to sentence complexity metrics such as Surprisal (Hale, 2001), which give a mathematical specification of the probability of the next word in the sentence.

Although correlations are found between the log-probability of the next-word and empirical sentence processing measures such as eye-tracking and self-paced reading times, complexity metrical accounts abstract away from a clear specification of a particular parsing strategy. Hale (2014) therefore proposed to relate the Chunking Theory of Learning (CTL) (Rosenbloom & Newell, 1987) to Surprisal Theory, with chunking as a mechanism that operates in accordance with this surprisal metric. Applying CTL to sentence processing, the idea is that parsing operators can be fused into a quicker executing macro-operator for more familiar sentence structures. Hale predicted that surprisal effects, signaling a higher processing difficulty, would occur exactly on those points where fused macro-operators cannot account for less common sentence structures.

In this study, Hale's application of the Chunking Theory of Learning (CTL) has been evaluated and examined. In section 1, a history of sentence comprehension theories will be presented and the predictions of these will be evaluated according to empirical findings on sentence processing difficulties. Furthermore, Hale's application of CTL will be considered in detail, leading to a substantiation of the relevance of our research statement in section 2. An improvement on Hale's method to test the predictions of CTL will be proposed in section 3, and the results of my execution of this method will be displayed in section 4 and discussed in sections 5 and 6.

1 Theoretical Background

1.1 Context-free grammars

Table 1

Example of a Context-Free Grammar

S → NP VP	DT → the
NP → NP VP	NN → horse
NP → DT NN	NN → barn
VP → VBD PP	VBN → raced
VP → VBD	VBD → raced
VP → VBN PP	VBD → fell
PP → IN PP	IN → past

Before discussing sentence parsing strategies, it is first necessary to have a detailed and well-defined theoretical account for the syntactic analysis of sentences. This article will reason from context-free grammars (CFGs) only, which render a rather simple mathematical model for sentence structure. In this formal system, rules are given such that single non-terminals can be rewritten as a sequence of terminals and/or non-terminals, providing a hierarchical system of sentence structure, while the rule choice does not depend on the context of these non-terminals (hence such a grammar is called context-free). Parsing itself is the process of constructing a syntactic analysis for a sentence; for CFGs this process can be defined as assigning the correct derivation to a sequence of terminals. CFGs can provide an account for various linguistic properties, such as syntactic types, hierarchical structure, constituency, syntactic ambiguity, and precedence relationships, but they are not able to express displacement phenomena, such as relative clauses, topicalization and questions (Hale, 2014, pp. 11-17). An example of a context-free grammar is displayed in table 1 above.

(1.1) *The horse raced past the barn fell.*

The famous sentence from Bever (1970) in example 1.1, which can be constructed by this grammar, shows that grammatical sentences might be ambiguous locally (having multiple possible analyses in the parsing process), while having only one possible derivation globally, as presented in figure 1. When the local ambiguity is resolved towards the dispreferred structure, the processing difficulties involved are referred to as the garden-path effect, because listeners or readers are then led into the incorrect derivation, figuratively having to trace back to be able to find the correct derivation. Being

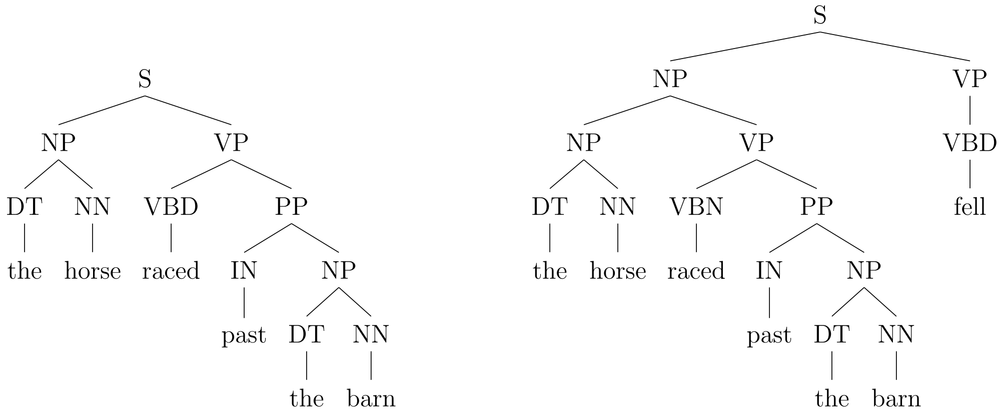

Figure 1. A local and global derivation for the sentence in example 1.1.

led into a garden-path is viewed as choosing the wrong structural alternative, and this abstract idea of garden-pathing has been formalized throughout the past decades.

1.2 Cognitive architecture

Besides choosing a formal grammar to reason from, it is important to determine which properties would nominate a parsing theory as a good candidate for modeling human parsing. That is, a good theory would reflect the powers and limits of the human mind in all its facets. The just explained effect of local ambiguity and selecting the incorrect derivation is only one of the properties of language that need a proper formalization. Crocker (1999) has mentioned three important properties of a cognitive model. Firstly, one basic property that a cognitive model should exhibit is *incrementality*, which is defined as the item-by-item (or in this case, word-by-word) processing of information. Secondly, the parsing mechanism should reflect how structural ambiguities (either local or global) are dealt with by a human. Thirdly, the processing complexity, for instance measured in time or space complexity, should increase for sentences which are shown to be more difficult by psycholinguistic experiments on sentence processing (Crocker, 1999). Such necessary properties are indicators of a good mental model of parsing, and should therefore be taken into account when the respective models are evaluated and compared.

1.3 Sentence parsing strategies

Some famous parsing mechanisms for context-free languages are top-down, bottom-up and left-corner parsing. Hale (2014) introduced these mechanisms by making use of pushdown automata.

In short, a top-down parser starts by assuming the word sequence is of type S (sentence) and works its way down the derivation tree by using depth-first search in expanding nodes according to the grammar rules. This algorithm assumes a hypothesis without checking it against the sentence, with the effect that at many points more than one rewrite rule can be chosen and there is no strategy to chose one above the other, resulting in a high degree of non-determinism.

Bottom-up parsing works in the opposite direction, starting from the encountered words by shifting them onto the stack, reducing combinations of words into higher categories according to the rewrite rules, and ending with the symbol S on the stack if the sentence is grammatical. In more detail, the two actions 'shift' and 'reduce' can be formalized as follows (Hale, 2014):

SHIFT: If the next word w in the sentence is a terminal in the grammar, push w onto the stack.

REDUCE: If the top of the stack contains a sequence of symbols $s_1 s_2 ... s_n$ and there is a grammar rule $X \longrightarrow s_1 s_2 ... s_n$, then pop the symbols from the stack and push X onto the stack.

This algorithm might seem more efficient, because unlike a top-down parser it does not work out a hypothesis about the sentence structure before having seen the words. However, similar to top-down parsing, it does operate with some degree of non-determinacy, since at some points more than one reduction rule might be applicable.

A more "psychologically plausible" algorithm, as Crocker (1999, p. 15) formulates it, would be left-corner parsing, which is a combination of top-down and bottom-up parsing. In this parsing strategy, the left-corner of a grammar rule is chosen to project the mother category of a (bottom-up) encountered word and predict the remaining categories on its right-hand side (top-down). This strategy captures some degree of expectation and anticipation in the human mind. A disadvantage of this algorithm is that it is less applicable to head-final languages than to head-initial languages, because heads are predictive for the rest of the phrase. A more universal mechanism would be Generalized Left-Corner (GLC) parsing, in which the left-corner can be stretched over more than one word (Demers, 1977). In

Table 2

A demonstration of how announce points are placed; every symbol left of the announce point is parsed bottom-up and predictive for the rest of the phrase, every symbol right of it is parsed top-down.

Announce point location	Symbols parsed bottom-up	Symbols parsed top-down	Example phrase
PP ⟶ P _ NP	P	NP	[on] [the beach]
VP ⟶ VP AdvP _	VP, AdvP		[sleep][furiously]
S ⟶ _ AdvP S		AdvP, S	[Fortunately][he agreed]

Note: Announce point locations are displayed as an underscore

GLC parsing, each rewrite rule receives an announce point on its right-hand side. The symbols left of the announce point are said to be predictive for the rest of the phrase. Depending on the placement of the announce point, bottom-up or top-down parsing will take place. Each rule is parsed bottom-up, until the announce point is reached. Then the predicted remainder of the rule is pushed onto the stack and parsed top-down. It is different from regular left-corner parsing only in its arbitrary placement of the announce point: in regular left-corner parsing, the announce point is always placed next to the first symbol on the rule's right-hand side. Table 2 shows for different grammar rules how the announce points are placed and which symbols are parsed bottom-up and which are parsed top-down. (Crocker, 1999; Demers, 1977; Hale, 2014).[1]

1.4 Garden Path Theory

The most well-known theory on the garden-path effect is the Garden Path Theory of Frazier (1978). This theory gives a heuristic account for sentence processing difficulties in parsing garden-path sentences: in a case of local ambiguity, ambiguity-resolution heuristics, such as the famous Minimal Attachment and Late Closure, are used to guide the parser into one interpretation. Minimal Attachment, for example, prefers derivations with less nodes over derivations with more, attaching incoming symbols into the phrase marker already being constructed. This also applies to the sentence in example 1.1: at first, the left analysis in figure 1 for ”the horse raced past the barn” is favored over the right, because the right derivation takes an extra NP node, signaling incompleteness of the phrase. The garden-path effect occurs when heuristics guide the parser into a globally incorrect and syntactically impossible interpretation, and can be resolved by backtracking to the place in the sentence where the wrong analysis was chosen and pursuing an alternative

[1] For more detailed information on context-free languages, pushdown automata, and automaton parsing mechanisms, read Crocker (1999), Hale (2014) and Sipser (1996).

analysis (Crocker, 1999). Some fundamental ideas of Garden Path Theory are the assumption that comprehension is single-path (as the words are heard or read, only one analysis is considered at a time), the principle of local indeterminacy and the modular approach that there are two subsequent stages in comprehension, namely a syntactic and a semantic stage, uninfluenced by each other (Hale, 2014). One might, however, doubt whether these stages are completely separate, because it has been found in many empirical studies that syntactic comprehension is influenced by factors such as prosody, referential context, animacy and thematic role assignment, favoring an approach of interaction between different linguistic modules (Crocker, 1999).

1.5 Sausage Machine Theory

Another theory that reasons from two separate stages is the Sausage Machine Theory by Frazier and Fodor (1978). In contrast to Garden Path theory, this theory does not assume modularity of different linguistic domains, but instead it states that there are two distinct steps of assigning syntactic structure to the word string. In the first stage, substrings of the sentence are assigned lexical and phrasal nodes by what is called the Preliminary Phrase Packager (PPP, or Sausage Machine), and in the second stage, these strings are linked together with higher nodes to combine them into a sentence by the Sentence Structure Supervisor (SSS). These two mechanisms have a considerably different behavior: the shortsighted PPP analyzes only a few words at a time, joining them into clauses or sub-clausal phrases, while the SSS can keep track of long-distance dependencies and commitments between phrases and takes into account the well-formedness rules when combining the phrases into higher non-terminal nodes.

One general motivation that Frazier and Fodor give for their model is that it considers the limits of the human cognitive architecture: in a single-stage parser, the computational complexity would increase exponentially with the sentence length, yet it is believed that amount of words is not a good predictor of sentence processing difficulty. In the two-stage Sausage Machine model, the computational complexity would not increase exponentially with sentence length, but the demand on working memory can be kept within reasonable limits, because analyses of subphrases can be dropped from the first stage as they are established in the second stage: the more structured the information that has to be stored, the lower the demand on working memory storage (Frazier & Fodor, 1978).

1.6 Surprisal Theory

One method to relate proposed parsing algorithms to corresponding observed difficulty measures in psycholinguistic experiments (e.g. reading times, eye-tracking or brain activity experiments) could be to calculate the correlation between the number of parsing operations and the observed difficulty. However, in a different perspective, one may start from information-theoretical accounts about processing difficulty and then relate them to real parsing mechanisms that might account for them (Hale, 2001, 2014, 2016). The intention is that processing difficulty on a word is high when its conditional probability is low, which means it has a low probability given the words already heard or read. This probability of a derivation can be calculated with use of probabilistic grammars, by determining the conditional probability of rewrite rules in a corpus, and then multiplying the probability of all rules that were applied in the given derivation.

One sentence complexity metric in the information-theoretical perspective is Surprisal, which calculates, given an incoming successor word w, the logarithm of the reciprocal of its probability as follows:

$$\text{surprisal}(x) = \log_2 \left(\frac{1}{PrefP(w)} \right) \tag{1}$$

In this equation, $PrefP(w)$ is the prefix probability at word w divided by the prefix probability at the previous word, or in simpler words, the change in probability of the previous derivation in contrast to the derivation when word w is encountered (Hale, 2014, 2016). Although Surprisal Theory is supported by several empirical studies , it models syntactic processing only computationally, not algorithmically: it gives a specification of how high the difficulty of parsing would be, given the syntactic structure of the sentence, but does not characterize a mechanism that operates in accordance with these metrics. One reason for the productivity of such complexity metrics is, in Hale's words, "the combinality of these information-theoretical complexity metrics with essentially any model of language" (Hale, 2014, p. 86), and that is exactly what he started working on.

1.7 The Chunking Theory of Learning

Hale suggested that surprisal could be viewed as a consequence of the Chunking Theory of Learning (CTL) (Rosenbloom & Newell, 1987), which states that cognitive operators can be fused together into a quicker executing macro-operator if used often. Applying this theory to sentence processing, the idea is that parsing operators such as 'shift' and 'reduce' (see the paragraph on bottom-up parsing in section 1.3) can be merged into one macro-operator if the sentence structure is more familiar: the operators are then learned to be a chunk. Hale's prediction was that surprisal effects would occur exactly on those points where highly general chunking mechanisms cannot account for syntactic structures that are less familiar for that particular context, which is among others the case for garden path sentences. A combination of CTL with Generalized Left-Corner parsing could give an explanation of how frequency and probability influence parsing. In order to identify potential chunks, he introduced *cohesion* as a measure of how well triples of parsing operators go together, such that a higher cohesion degree would signal they are more likely to become a macro-operator. This cohesion value can be calculated, based on Manning & Schütze (1999), as the likelihood-ratio between the (null) hypothesis that parsing actions are probabilistically independent of previous parsing actions, and the hypothesis that they *are* dependent. A high cohesion value for one action triple would signify that this triple is more likely under the hypothesis that a parsing action would follow the two previous parsing actions than its base rate of occurrence would suggest, therefore being more likely to become a macro-operator under CTL. This will be formalized further in section 3. Using linear regression to predict eye-fixation duration in an English and French eye-tracking corpus, Hale showed that cohesion degree of chunks is indeed a positive predictor of comprehension (Hale, 2014, ch. 8), supporting the idea that parsing action chunks get stronger with usage.

2 Research Statement

The purpose of this study is to investigate further Hale's claim that chunks of parsing operators can be learned to go together, resulting in faster parsing for familiar sentence structures, while inducing surprisal effects for less familiar structures. In his study on English, Hale used the Charniak parser (Charniak & Johnson, 2005), which is an automatic, and therefore slightly flawed, parser, to obtain the phrase structures for the prominent English eye-tracking corpus called Dundee Corpus (Kennedy & Pynte, 2005). To improve on the validity of this method, we will evaluate on chunking by using the Natural Stories Corpus (Futrell et al., 2018), which is an already parsed and hand-corrected reading-times corpus of English: the fact that it is hand-corrected will prevent that wrongly parsed sentences are taken into account in the calculations, in contrast with what would be the case for an automatic parser. As our research statement, we will argue that our results are in agreement with the prediction that parsing action chunks get stronger with usage, resulting in faster reading times for more familiar parsing action sequences, while inducing surprisal effects and longer reading times for less familiar parsing action sequences. In the following sections, we will give a thorough explanation of our methodology, our results and their implications.

3 Method

In the same way as Hale went through the corpora to come up with a list of candidate chunks ranked by cohesion degree, and similarly to how he compared these measures to the eye-tracking data, we have analyzed the data of the Natural Stories Corpus (Futrell et al., 2018). This corpus consists of English texts, containing many low-frequency sentence structures, without any disfluency effects for native speakers. It is annotated with hand-corrected parse trees in Penn Treebank-style and includes self-paced reading (SPR) times, averaged out over 181 native English speakers. Firstly, we have calculated the cohesion of all parsing action bigrams and trigrams, by extracting from the corpus the frequencies of the individual occurrences of parsing actions and how often they appear together, using the association measure as introduced by Manning and Schütze (1999). Secondly, we have analyzed the correlation between cohesion degree and reported self-paced reading times, averaged over participants (mean RTs).

As for the calculation of cohesion, this is a likelihood-ratio between the hypothesis that parsing actions are probabilistically independent of previous parsing actions and the hypothesis that they *are* dependent. Formally, these alternative explanations for the occurrence of a trigram of parsing actions $a_1a_2a_3$ can be represented as follows, where hypothesis 1 is a formalization of independence and hypothesis 2 a formalization of dependence:

$$\begin{array}{l} \textbf{Hypothesis 1. } P(a_3|a_1a_2) = p = P(a_3|\neg(a_1a_2)) \\ \textbf{Hypothesis 2. } P(a_3|a_1a_2) = p_1 \neq p_2 = P(a_3|\neg(a_1a_2)) \end{array} \tag{2}$$

The likelihood of occurrences of a_1a_2, a_3, and $a_1a_2a_3$ is calculated as in equation 3 for the two hypotheses, where b is the binomial probability mass function, $b(k,n,p) = \binom{n}{k}p^k(1-p)^{n-k}$, where c stands for the frequency count of an action or action tuple, $p = \frac{c_3}{N}$, $p_1 = \frac{c_{123}}{c_{12}}$, $p_2 = \frac{c_3 - c_{123}}{N - c_{12}}$, and N is the total amount of parsing actions in the corpus.

$$\begin{aligned} L(H_1) = {} & b(c_{123}, c_{12}, p) \cdot b(c_3 - c_{123}, N - c_{12}, p) \\ L(H_2) = {} & b(c_{123}, c_{12}, p_1) \cdot b(c_3 - c_{123}, N - c_{12}, p_2) \end{aligned} \tag{3}$$

The logarithm of the likelihood ratio is then given by the following equation:

$$\begin{aligned}\log\lambda &= \log\frac{L(H_1)}{L(H_2)}\\ &= \log\frac{b(c_{123},c_{12},p)\cdot b(c_3-c_{123},N-c_{12},p)}{b(c_{123},c_{12},p_1)\cdot b(c_3-c_{123},N-c_{12},p_2)}\\ &= \log L(c_{123},c_{12},p)+\log L(c_3-c_{123},N-c_{12},p)\\ &\quad-\log L(c_{123},c_{12},p_1)-\log L(c_3-c_{123},N-c_{12},p_2)\end{aligned} \tag{4}$$

where $L(k,n,p) = p^k(1-p)^{n-k}$. By convention, the cohesion value is displayed as the log likelihood ratio multiplied by negative two, resulting in the quantity $-2\log\lambda$ (Manning & Schütze, 1999)[2].

To come up with a list of cohesion values per word, in order to be able to find a correlation between cohesion degree and mean RT per word, we have calculated the maximum and mean of the bigram and trigram cohesion values for the parsing actions for each word (this is because words often would be considered by more than one parsing action). In addition to that, for each word we have taken a look at the parsing action trigram whose middle action was the shift of that word; this is what Hale called the ”presumably relevant chunk” (Hale, 2014, p. 95) and the cohesion value that he used in his study. The final step was to calculate the Pearson correlation between the calculated cohesion values and the mean RT per word. If the hypothesis that chunks of parsing operators can be learned to go together, resulting in faster parsing for more familiar sentence structures, is true, we would expect a negative correlation between cohesion degree of parsing action chunks and mean RT: the higher the cohesion of parsing actions on a word, the shorter the reading time expected.

4 Results

Table 3 presents some parsing action triples with corresponding frequencies as counted in the corpus, and their calculated cohesion values (note that the bar after a phrase marker indicates that the phrase is not ended yet). The Pearson correlation coefficients for the different calculated cohesion values per word (maximum, mean and cohesion of the relevant parsing action trigram as in Hale (2014), see section 3) and the mean RT on that word in the corpus are presented in table 4. The results show a significant but small negative

[2] The equations for pairs of actions, taken from and substantiated by Manning & Schütze (1999, §5.3.4), are customized here for triples of actions.

Table 3

Some Example Parsing Action Trigrams from the Natural Stories Corpus with Frequencies and Cohesion Values

$a_1; a_2; a_3$	c_1	c_2	c_3	c_{12}	c_{123}	$-2 \log \lambda$

reduce-binary VP;
reduce-unary I-BAR;
reduce-binary IP
	2161	1531	1531	999	893	5587.1601

⋮	⋮	⋮	⋮	⋮	⋮	⋮

shift;
reduce-unary VP;
reduce-unary VP-BAR
	11727	118	2161	112	46	114.8011

⋮	⋮	⋮	⋮	⋮	⋮	⋮

reduce-binary IP;
shift;
reduce-unary AdvP
	1513	11727	325	266	1	1.1075

correlation between cohesion degree of action trigrams and mean RT (only for the maximum trigram cohesion and cohesion of the relevant trigram). This means, participants spent significantly less time on the words with a higher maximum trigram cohesion and a higher cohesion of the relevant trigram (the parsing action triple with as second action the shift of that word). An unexpected positive correlation between the average cohesion value of all bigram actions for a word and the average reading time on that word was found.

Table 4

Pearson Correlation Coefficient for the potential Predictors of the Mean RT, together with Significance Value

Predictor	Pearson's r	2-tailed p-value
Maximum bigram cohesion	.013	.169
Mean bigram cohesion	.043	.008*
Maximum trigram cohesion	−.027	.006*
Mean trigram cohesion	−.011	.271
Cohesion of the relevant trigram	−.023	.021*

*$p < .05$

5 Discussion

The results are in agreement with Hale's results (Hale, 2014) and in line with our expectations: a significant negative correlation is found between cohesion degree of parsing operator trigrams and average reading times, in accordance with Hale's analysis that the degree to which triples cohered was a negative predictor of the average eye-fixation duration in English.

It is, however, hard to make a strict comparison between the two studies. Firstly, because the data of the corpora could not be directly compared: the Dundee corpus, as used by Hale, applied a different reading complexity metric than the Natural Stories Corpus, as used in the current study, namely eye-fixation duration versus self-paced reading times. Moreover, the corpora made use of different reading genres, namely newspaper articles versus written stories. A second reason why a strict comparison between the studies is hard to make, is the fact that the Natural Stories corpus was annotated with a different parser than Hale used on the Dundee corpus, namely the hand-corrected Penn Treebank-style parse trees versus the automatic Charniak parser. A third reason is that we used a different method of analyzing the relation between chunking and reading difficulty: as opposed to Hale, who used a linear regression analysis model, we have used the Pearson correlation measure. It is encouraging to find that despite these discrepancies, our results are in agreement with the results of Hale. This might therefore indicate a generalization of the effect over different corpora, different processing difficulty measures and different parsers.

As for the significant correlation which was found for the mean bigram cohesion of all parsing actions for a word and its the mean RT, this correlation was positive, which means the higher the average cohesion of all bigram actions for a particular word, the longer the reading duration averaged over participants. Even though this correlation is only small, we could find no other explanation for this unexpected result than that parsing action bigrams could be too small to be learned a chunk, such that a calculation of a cohesion value for bigrams is relatively meaningless.

Furthermore, it is important to mention that there are other predictors of mean RT which we have not accounted for in this short study. Futrell and colleagues (2018, p. 78) reported that basic psycholinguistic effects are present in the SPR data of the Natural Stories Corpus: frequent words were read faster, longer words were read more slowly and words with a higher surprisal (having a lower log probability under a word trigram model) were read more slowly. In addition to these, other effects could be thought of, such as the effect that words at the ends of sentences are read more slowly, which is the so-called sentence wrap-up effect (Just & Carpenter, 1980), and the processing delay effect called spill-over, in which increased processing (slower reading) does not take place on the word causing the increase, but on one or two words later in the sentence (Just, Carpenter, & Woolley, 1982).

Based on the found agreement in results between Hale's study and the current, we conclude that the suggestion that parsing actions can be learned by practice is a good start of a promising approach on sentence processing, which in future studies should be examined in more detail by considering all possible effects at play, and accounting for the necessary properties of a parsing model. In comparison to the Garden Path Theory and Sausage Machine Theory, the Chunking Theory of Learning offers a more direct link to sentence processing measures, because log likelihood ratios of parsing action chunks can be directly compared to processing difficulty measures. One negative implication of the Garden Path Theory (Frazier & Rayner, 1982) is that it reasons from a separate syntactic and semantic stage, while empirical evidence advocates interaction between these linguistic domains (Crocker, 1999; Tanenhaus, Spivey-Knowlton, & Hanna, 2000). Besides that, studies on the theory that there is a preference ordering on automaton parsing actions have shown that there is no fixed set of heuristic principles that performs well enough to be taken as a psycholinguistic a law. Chunking, however, could offer a reasonable and psycho-linguistically plausible explanation for garden path effects. Compared to the Sausage Machine approach of Frazier and Fodor (1978), CTL might be more practical to apply to different parsers and therefore can be more thoroughly examined in empirical research. Maybe it could, for example, be applied to parsing techniques for more complex grammars (e.g. using complex categories or lexicalized grammars to express displacement phenomena), thereby approaching a natural language theory.

6 Conclusion

In this study, we have examined Hale's proposal (Hale, 2014) that the Chunking Theory of Learning can be considered a concrete mechanism to relate to the sentence complexity metric of Surprisal, such that sentence parsing operators can be fused into a quicker macro-operator if practiced often, accounting for the comprehension effects as found for different sentence structures. We have tested his predictions on the Natural Stories Corpus (Futrell et al., 2018), which is a self-paced reading corpus, annotated with hand-corrected parse trees. The results demonstrate a correlation between cohesion degree of parsing operator trigrams and average reading times, supporting the statement that parsing action triples can be learned to be a chunk. Therefore, we conclude that the presented results are in line with Hale's results and are in support of his predictions: it is likely that triples of parsing operators can be learned to be a chunk when having a higher cohesion, such that surprisal effects are reduced for more familiar sentence structures.

From the above evaluation and reflection (section 5), we further conclude that more research is needed to be able to give an answer to the risen questions about differences between corpora and analyses, external effects and whether the implications of the proposed theories are in agreement with empirical findings on human sentence processing and the human sentence parsing mechanism.

References

Bever, T. G. (1970). The cognitive basis for linguistic structures. *Cognition and the development of language*, *279*(362), 1–61.

Charniak, E., & Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 173–180).

Crocker, M. W. (1999). Mechanisms for sentence processing. *Language processing*, 191–232.

Demers, A. J. (1977). Generalized left-corner parsing. In *Proceedings of the 4th acm sigact-sigplan symposium on principles of programming languages* (pp. 170–182).

Frazier, L. (1978). On comprehending sentences: Syntactic parsing strategies. *Doctoral Dissertation.*

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, *14*(2), 178–210.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2018). The natural stories corpus. In *Proceedings of lrec 2018, eleventh international conference on language resources and evaluation* (pp. 76–82). Miyazaki, Japan. Retrieved from `https://github.com/languageMIT/naturalstories`

Hale, J. T. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the association for computational linguistics on language technologies* (pp. 1–8).

Hale, J. T. (2014). *Automaton theories of human sentence comprehension.* Center for the Study of Language and Information.

Hale, J. T. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397–412.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of experimental psychology: General*, *111*(2), 228.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision research*, *45*(2), 153–168.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT press.

Rosenbloom, P. S., & Newell, A. (1987). Learning by chunking: A production-system model of practice. *Production System Models of Learning and Development*, 221–286.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Sipser, M. (1996). Introduction to the theory of computation. *ACM Sigact News*, *27*(1), 27–29.

Tanenhaus, M. K., Spivey-Knowlton, M. J., & Hanna, J. E. (2000). Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. *Architectures and mechanisms for language processing*, 90–118.