Surprise Effects in Language as a Consequence of Chunking

John Hale
Department of Linguistics
Cornell University
USA

Stéphane Rauzy and Philippe Blache
Brain and Language Research Institute
Aix-Marseille University
France

Author Note

The authors are grateful to Stefan Frank and Alan Kennedy for help with the Dundee corpus. This research was supported by NSF CAREER award number 0741666
Abstract

The Chunking Theory of Learning, originally proposed within the context of work on cognitive architecture, provides a mechanism-level explanation for ubiquitous surprise effects in human sentence processing. To flesh out this connection, we simulate the effect of chunking in a parsing task. Considering French and English, this yields an inventory of chunks, which can be interpreted both procedurally and with respect to a grammar. Entering the cohesion of these chunks into a regression against human eye fixation durations demonstrates that chunks, as parsing macro-actions, could indeed relate familiarity and speedup in the way that the descriptive surprisal correlation suggests.
Surprise Effects in Language as a Consequence of Chunking

Introduction

This theoretical squib connects two proposals that are relatively well-established within cognitive science. One is the Chunking Theory of Learning, Rosenbloom and Newell’s account of practice curves (1987). The other proposal is that the speed of language-processing should be related to the (log) conditional probability of a word in its left-context. This idea, dubbed ‘surprisal’ in Hale (2001) has been applied widely in psycholinguistics with a variety of probability models (Boston, Hale, Kluegl, Patil, & Vasisht, 2008; Demberg & Keller, 2008; Levy, 2008). We can understand surprisal at Marr’s (1982) most abstract level as the claim that a processor should be sensitive to probability distributions in its language environment; namely conditional probabilities given the left context. If we fix on a parsing mechanism of some sort, then combining it with chunking yields a mechanistic explanation for surprisal at Marr’s middle level, the Algorithm and Representation level.

An important line of work takes a surface-oriented view, effectively treating chunks in language as word-collocations (Hammerton, Osborne, Armstrong, & Daelemans, 2002). By labeling our chunks with sequences of parser actions themselves, we break with this tradition. Such a break is significant in the cognitive science of language because it offers a way to reconcile several long-standing oppositions. One of these is between what Michael Tanenhaus has called the ‘language as product’ tradition as opposed to the ‘language as action’ tradition. Chunks are meaningful from both perspectives. On the one hand, as the output of a process like production compilation, they are bona fide mental products. But on the other, they are still procedural knowledge — knowledge of how to do an action. Another longstanding opposition is between formalism and functionalism. These are rival philosophies of linguistics that prioritize different research questions and strive towards different sorts of explanations. Here too, the chunking viewpoint offers a possible rapprochement. With Bybee (2010, 34) we can say that constructions get stronger
with usage, while at the same time interpreting them with respect to particular rules of a generative grammar as envisaged e.g. by Hornstein (2009, 146).

We present two corpus studies, one in French and one in English. Both deal with syntactic parsing, using a common formalization of chunking. Both show that this formalization derives the sort of speed up that has been the hallmark of surprisal. The results confirm Abney’s 1991 insight that function words and content word ought to be parsed together as a unit. The squib concludes with a survey of related work.

**Incremental parsing mechanism**

In order to do any chunking, there must be some basic operators available for the task. For perceiving sentence structure, the simplest proposal is that analysis operators map one-to-one on to grammar rules. As E. Stabler (1991) argues, any other formulation takes on an obligation to explain the mismatch. Following Johnson-Laird (1983) and others we might suppose a left-corner parser as shown in Figure 1.

Figure 1(b) shows the stream of left-corner parser actions in a traversal of the tree in 1(a). The operators labeled ‘shift’ are the ones that deal with words as opposed to syntactic structure. These shifts come in the same linear order as the words of the sentence; the mechanism is an *incremental* parser. Of course, a full model will also require an oracle for resolving cases where more than one action is applicable; see Hale (2011) for one approach. In what follows, we re-use the same stack-based parsing strategy as in that paper. It is a case of Generalized Left Corner parsing (Demers, 1977) whose operations can be abstractly characterized into three schemas:

project  if the top of the stack is a symbol $Y$, and there is a grammar rule $X \rightarrow Y \beta$ whose right-hand side starts with $Y$, then replace $Y$ with two new symbols: a record that $X$ has been found, and an expectation for each of the remaining right-hand side symbols e.g. $[Z_1], [Z_2], [Z_3], \ldots$ where $\beta = Z_1 Z_2 Z_3 \ldots$
Figure 1. Traversal according to a left-corner parsing strategy. Bold lines name parser actions, non-bold lines show stack contents. Stacks grow to the left and expectations are identified by square brackets.
project+complete if the top of the stack is \( Y \), and right below it is an expectation \([X]\), then replace both with the new expectations \([Z_1]\), \([Z_2]\), \([Z_3]\)… there by eliminating \([X]\).

shift if the next word of the sentence is one of the grammar’s terminal symbols, push \( w \) on to the top of the stack symbolizing that word.

The particular strategy that we examine in this paper is mostly left-corner, but bottom-up on “adjunction” rules like \( \text{NP} \rightarrow \text{NP PP} \). Inspired by work such as Clifton, Speer, and Abney (1991) and Hemforth, Konieczny, and Strube (1993), this mixed strategy avoids predicting optional postmodifiers.

**Chunking without iteration**

Chunking, in the sense that is relevant here, refers to the creation of macro-operators from sequences of ordinary task operators. In the most well-known formulation, production compilation, there is a small but non-negligible probability that any two operators may fuse together. This leads iteratively to larger and larger macro-operators, but only in cases where the two successive operators co-occur frequently.

Instead of developing the collection of macro-operators one by one, it is possible to simply tabulate sub-sequences of operator names that tend to cohere. The basic idea is to view streams of parser actions, like the one in Figure 1(b), as a new text. In this new text the names of the parser actions are each individual words. Then any method for finding collocations may be used. Table 1 presents a representative selection of examples. In this table, the column labelled **cohesion** gives a log-likelihood ratio between the expected value if the individual parsing actions were (probabilistically) independent as compared to the observed number of attestations. This measure of association is introduced pedagogically in Manning and Schutze (1999, §5.3.4) using sequences of size 2, but the idea generalizes straightforwardly to sequence of any size. The chunks of size 3 in the tables below were obtained using Pedersen’s N-Gram Statistics Package (Banerjee & Pedersen, 2003) to traversals of the Penn Treebank’s Wall Street Journal sub-corpus (Marcus, Santorini, &
Marcinkiewicz, 1993) according to the mostly left-corner strategy introduced above in the section entitled “Incremental parsing mechanism.”

<table>
<thead>
<tr>
<th>chunk</th>
<th>cohesion</th>
</tr>
</thead>
<tbody>
<tr>
<td>shift preposition ; project PP → P NP ; shift determiner</td>
<td>472371.7530</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>shift determiner ; project NP → Det N ; complete N</td>
<td>349337.5764</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>project NP → Det Adj N ; complete Adj ; complete N</td>
<td>144445.6793</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>project NP → Adj ; shift colon ; shift verb</td>
<td>10.0435</td>
</tr>
</tbody>
</table>

Table 1

Example chunks from an English Treebank

A similar pattern emerges when the same analysis is applied to the FTB_{LPL}, a version of the French Treebank (Abeillé, Clément, & Toussenel, 2003) that has been adjusted in various ways to better conform to the Penn annotation style (Blache & Rauzy, 2012; Schluter & van Genabith, 2007). In the same manner as Table 1, Table 2 shows a selection of highly cohesive parser-action chunks, as well as one low-ranking, uncohesive chunk to illustrate the range of values taken by the particular measure of association, the log-likelihood ratio.

Empirically-derived chunks

Setting aside punctuation, the extracted chunks largely reflect Abney’s 1991 proposal. Abney suggested that parser actions should be chunked into units that include (roughly) a function word, the content word selected by that function word, and any other content words that linearly intervene between the two. We see this already in rules like NP → Det Adj N (English) and project NP → Det Noun AP (French) where both a noun
SURPRISAL AND CHUNKING

<table>
<thead>
<tr>
<th>chunk</th>
<th>cohesion</th>
</tr>
</thead>
<tbody>
<tr>
<td>shift preposition ; project PP → Prep NP ; shift determiner</td>
<td>101154.2114</td>
</tr>
<tr>
<td></td>
<td>:</td>
</tr>
<tr>
<td>shift determiner ; project NP → Det Noun ; complete N</td>
<td>100271.7340</td>
</tr>
<tr>
<td></td>
<td>:</td>
</tr>
<tr>
<td>shift determiner ; project NP → Det Noun AP ; complete N</td>
<td>79041.4928</td>
</tr>
<tr>
<td></td>
<td>:</td>
</tr>
<tr>
<td>project NP → Adj ; shift colon ; shift verb</td>
<td>10.0435</td>
</tr>
</tbody>
</table>

Table 2

*Example chunks from a French Treebank*

and an adjective form part of a noun phrase. Both elements are incorporated as part of a prediction that can be made upon encountering the phrase-initial determiner.

**Reading faster when the chunk is familiar**

In their Chunking Theory of Learning, Rosenbloom and Newell (1987) hypothesize a relationship between chunks and processing time:

**performance assumption** The performance program of the system is coded in terms of high-level chunks, with the time to process a chunk being less than the time to process its constituent chunks.

They also make a proposal about how chunks are learned:

**learning assumption** Chunks are learned at a constant rate on average from the relevant patterns of stimuli and responses that occur in the specific environments experienced.

In the case of sentence parsing, these environments are naturally viewed as pieces of tree structure — or rather instructions to build that tree structure as in Figure 1(b). For Newell and Rosenbloom, a chunk is either in the mind or not. However we can take the
log-likelihood as an indication of how likely an agent using something like
production compilation would be to have acquired this chunk. Having formed the chunk,
the agent can use it, rather than executing each of its component structure-building actions
separately, and thereby accomplishing the same amount of parsing work in less time. This
follows by the performance assumption.

To evaluate this idea, we used linear regression to predict eye-fixation durations. In
fitted models for both French and English, the degree of cohesion was a significant negative
predictor of the human participants' fixation duration. That is, people read more quickly
when the word they were looking at would be part of a highly cohesive chunk. This effect
obtained even when other factors such as lexical frequency and number of characters were
entered as co-predictors\(^1\). While this study considers only unlexicalized chunks of size 3, we
should expect the basic equation between chunk-cohesion and reading speed to hold more
generally since larger chunks are necessarily less frequent.

**English study**

We used a version of the Dundee eyetracking corpus (Kennedy & Pynte, 2005) that
was prepared by Stefan Frank (2011). This sample of newspaper text is marked-up with
parts of speech tags using the Penn tag set. From these tags, we used the Charniak parser
to obtain phrase structures (Charniak & Johnson, 2005). Training separately on the
BLLIP corpus (1999) yielded a list of unlexicalized candidate chunks ranked by degree of
cohesion as discussed above in the section entitled "Chunking without iteration". This
training is appropriate since both chunks and the fixation durations come from the same
genre of writing. We considered just chunks of length 3 whose "middle" action was the
shift of the part of speech tag for fixated word. Table 3 reports a significant \(t\)-value for
scaled log-likelihoods as the only predictor of first-pass reading time.

Following Demberg and Keller (2008) we examined this correlation in combination

\(^1\)These co-predictors would be expected to be highly correlated with chunk cohesion on many accounts
of language change, and we find that they are.
with other predictors. Table 4 shows t-values for estimated coefficients when predicting an early measure, single fixation duration, along with other factors that are known to influence reading time. The dependent variable here is log-transformed. The results are consistent

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>sig at p &lt; 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5.2807527 0.0003774</td>
<td>13993.605</td>
</tr>
<tr>
<td>number of chars</td>
<td>0.0083617 0.0002149</td>
<td>38.905 yes</td>
</tr>
<tr>
<td>word prob</td>
<td>-0.0136493 0.0002259</td>
<td>-60.433 yes</td>
</tr>
<tr>
<td>cohesion</td>
<td>-0.0010350 0.0003840</td>
<td>-2.695 yes</td>
</tr>
</tbody>
</table>

Table 4

*English: with other predictors*

with other findings suggesting a strong influence of predictability in early measures but less so in late measures (Boston et al., 2008; Boston, Hale, Vasishth, & Kliegl, 2011). We were unable to find a significant effect of cohesion with any larger aggregated measure in this sample. However, this did obtain with the French data discussed in the next section.

**French study**

For the French study, we used the pilot eyetracking data of Rauzy and Blache (2012). This sample of newspaper text includes a wealth of linguistic annotation, including phrase structure and part-of-speech tags. This rich markup made parsing unnecessary. We obtained a chunk inventory from the $FTB_{LPL}$. The pilot eyetracking data comes from 13
readers reading a small section of this same text, 198 sentences out of 4741. Again choosing the “middle” action for chunks of length 3 we considered mean fixation duration on a word, across readers. The log likelihood ratio of the presumably relevant chunk emerged as a significant negative predictor of this average fixation duration.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>sig at p &lt; 0.01 ?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.0239</td>
<td>2.899</td>
</tr>
<tr>
<td>cohesion</td>
<td>-5.466 x 10^{-4}</td>
<td>7.762 x 10^{-5}</td>
</tr>
</tbody>
</table>

Table 5

French: as only predictor

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>sig? at p &lt; 0.01 ?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.2155</td>
<td>3.338</td>
</tr>
<tr>
<td>cohesion</td>
<td>-1.111 x 10^{-4}</td>
<td>4.365 x 10^{-5}</td>
</tr>
<tr>
<td>number of chars</td>
<td>21.22</td>
<td>0.4912</td>
</tr>
<tr>
<td>word prob</td>
<td>16.08</td>
<td>0.5794</td>
</tr>
</tbody>
</table>

Table 6

French: with other predictors

Discussion

Above the word level, the cohesion of empirically-derived chunks helps to explain fixation times in eyetracking. This finding is consistent with other modeling work that applies techniques from computational linguistics to the prediction of psychological variables in reading (Boston et al., 2008; Demberg & Keller, 2008; van Schijndel & Schuler, 2013). What is new is recognizing the connection to Rosenbloom and Newell’s Chunking Theory of Learning. If human practice really does reflect the fusion of operators into macro-operators as a result of temporal contiguity, as the theory suggests, then we
should see a frequency-sensitive speed up. But this is no more and no less than exactly what surprisal, in the sense of Hale (2001) asserts². The simulations reported here suggest that this candidate mechanism, which is independently motivated in other domains of cognition, can derive the surprisal correlation itself.

The Chunking Theory of Learning sets up an algebra for combining frequency-sensitive representations. It requires a base of initial elements; that is to say there must be some underlying theory of how parsing happens. Following the Competence Hypothesis (Bresnan, 1982; Chomsky, 1965) we might as well assume that this basis corresponds to individual rules of some grammar. These can be used in standard ways to define a parsing automaton which in turn induces a search space as in Abney (1991) or Hale (2011). Of course, the idea is generic and could be applied to any transition-based parsing system for any formalism: dependency grammar (Nivre, 2006), unification grammar (Ytrestål, 2013), tree-adjointing grammar (Demberg, Keller, & Koller, 2013) or minimalist grammars (E. P. Stabler, 2013). As chunked macro-operators become available, the granularity of the search space becomes coarser. Such a conception represents a step towards simultaneously accounting both for the frequency-sensitive character of linguistic performance, as well as our ability to generalize combinatorially to new structures when necessary. This creative ability would still be explained by appeal to the basic operations, though — rather than learned chunks.

Related Work

There has been tradition within computational linguistics of enlarging the domain of locality (Joshi, Levy, & Takahashi, 1975). However, these efforts have typically focused on tree structures, rather than traversals of them. For instance, Rayner, Carter, and

²It is perhaps unfortunate that the title of the 2001 paper mentions a specific algorithm for computing surprisal because in fact, it is a purely normative claim: processing difficulty ought to be proportional to the logarithm of the reciprocal of the probability of a word given its left-context. This claim should be understood at Marr’s most abstract level of analysis, what he termed the ‘computational’ level.
Samuelsson (2000) chunk unification grammar rules rather than particular actions in their incremental parsing algorithms. Sangati and Keller (2013) derive an incremental parser for tree-substitution grammar, but in that setting, the subtrees are already chunked as much as they are ever going to be. The same applies to super-tagging, which presuppose some set of elementary trees (Bangalore & Joshi, 2010). Neither approach acknowledges a mechanism by which syntactic analysis steps are combined into larger steps. It is this latter point that serves as the foundation for Rosenbloom and Newell’s general account of practice effects in human cognition.

**Conclusion: surprisal as a practice effect**

We can interpret surprisal’s correlation between (log) probability and reading time as a descriptive claim. Chunking offers a mechanistic explanation for this description. The explanation is that in cases where people are reading faster, they are using larger macro-operators that were previously formed by chunking on the basis of frequent use. This explanation harmonizes with the classic competence-performance distinction in generative grammar. Whereas grammar rules still might characterize the knowledge a person has of his or her language, those same rules only describe linguistic performance in cases where the person is generalizing to unfamiliar cases. For the bulk of everyday language processing, a coarser search over highly-cohesive chunks is involved.

At bottom, the proposed explanation reframes surprisal effects in language processing as practice effects. In real language use, people recognize sentence structure, and cohesion values model their degree of practice with this specific skill. With high-density longitudinal studies like the Human Speechome Project (Roy et al., 2006) it may be possible to discern more details of how chunking proceeds, perhaps even defining a specific threshold.
References


