Variable beam search for generative neural parsing and its relevance for the analysis of neuro-imaging signal

Benoît Crabbé
LLF - CNRS
University of Paris - IUF
Place Paul Ricoeur
75013 Paris France
benoit.crabbe@linguist.univ-paris-diderot.fr

Murielle Popa-Fabre
ALMANACH - INRIA - LLF
University of Paris
2 rue Simone Iff
75012 Paris France
murielle.fabre@inria.fr

Christophe Pallier
Cognitive Neuroimaging Lab
INSERM-CEA
Neurospin bat.145
Gif-sur-Yvette France
christophe@pallier.org

Abstract

This paper describes a method of variable beam size inference for Recurrent Neural Network Grammar (RNNG) by drawing inspiration from sequential Monte-Carlo methods such as particle filtering.

The paper studies the relevance of such methods for speeding up the computations of direct generative parsing for RNNG. But it also studies the potential cognitive interpretation of the underlying representations built by the search method (beam activity) through analysis of neuro-imaging signal through correlations with the neuro-imaging signal during naturalistic text listening

1 Introduction

The paper provides a preliminary investigation of how the improvements in natural language parsing techniques can be used for modelling brain activity during online sentence comprehension. Specifically, we focus on the RNNG generative parsing framework (Dyer et al., 2016), that allows us to extract information-theoretic measures that are relevant for cognitive studies such as surprisal or entropy (Hale, 2016).

RNNG is essentially a generative parsing framework but it is not straightforward to design a one-step generative parser. Dyer et al. (2016) achieves generative parsing with RNNG in two steps: by reranking a discriminative parser. Stern et al. (2017) points out that the main difficulty comes from lexical biases using naive beam search. They propose new methods for modifying a classical word-synchronized beam search for avoiding the lexical bias problem.

The paper presents a substantially different search method for parsing in one-step with a neural generative parser, proposing a method inspired by sequential Monte-Carlo sampling and particle filtering (Doucet and Johansen, 2011; Buys and Blunsom, 2015).

Crucially, the proposed method is naturally not sensitive to problems of lexical biases faced by standard beam search. As a result, the parser is framed as generative and strictly incremental (without lookahead) while remaining quite accurate and generally more efficient than a more traditional beam search method. Not only, the proposed search method is a variable-beam search method, but as such, it naturally provides an additional method to measure the activity required to make the parser progress incrementally and word-by-word that can, in principle, be used for cognitive modelling purposes.

Taking advantage of our search method properties, we introduce new parsing complexity measures such as beam-size and beam-activity, adding to the well known entropy or surprisal that are already provided by the generative model. Finally, we evaluate the potential cognitive relevance of such measures for analyzing neuro-imaging data. Instantiating syntactic processing in terms of parser's analyses has already highlighted the language brain network to a large extent in the neuroimaging literature, we therefore expect that the goodness of fit of our parsing model with fMRI signal can be taken as an index of cognitive relevance. In other words, we aim at using our parsers' beam activity as a *proxy* modelling the kind of processes that might take place during online sentence comprehension, and thus investigate the potential cross-fertilization between computational solutions and cognitive-neuro imaging approaches to human parsing. The paper is organized as follows. Section 2 first describes an in-order variant of RNNG that is used as underlying formal framework supporting our experiments. It describes the variable beam search method inspired by particle filtering and finally introduces metrics quantifying

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing, pages 1150–1160,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

online parsing process. Section 3 provides empirical measures and quantifies the behaviour of the parser and of the proposed search method. Section 4 describes the relationship between the parsing model and the analysis of neuro-imaging data, section 5 presents the results.

2 RNNG and language modelling

2.1 In-order RNNG

We use an in-order variant of RNNG (Liu and Zhang, 2017; Kuncoro et al., 2017) where the set of transitions echoes the one of a left-corner parser. A configuration is a triple $\langle \mathbf{S}, \mathbf{B}, n\rangle$ where $\mathbf{S}$ is a stack, $\mathbf{B}$ is a buffer and n a counter of open brackets.

Init $\langle \emptyset, x_1 \ldots x_n, 0\rangle$

Goal $\langle S_0\bullet, \emptyset, 0\rangle$

Generate(x) $\frac{\langle \mathbf{S}, \bullet x_i | \mathbf{B}, n\rangle}{\langle \mathbf{S} | x_i \bullet, \mathbf{B}, n\rangle}$

Open(N) $\frac{\langle \mathbf{S} | S_0 \bullet, \mathbf{B}, n\rangle}{\langle \mathbf{S} | \bullet N | S_0 \bullet, \mathbf{B}, n+1\rangle}$

Close $\frac{\langle \mathbf{S} | \bullet N_i | \ldots | S_0 \bullet, \mathbf{B}, n\rangle}{\langle \mathbf{S} | N \bullet, \mathbf{B}, n-1\rangle}$

Each **Open** action is parametrized by a nonterminal symbol, thus there are as many nonterminals in the set A of actions as there are nonterminal symbols. As the parser is generative, there are as many generate actions as there are words x in the vocabulary X. Note however that we constrain the generate action to be the words $x_1 \ldots x_n$ of the actual sentence to parse. Observe also that the set of actions differs from the standard top down formulation of RNNG (Dyer et al., 2016) and from the in-order transition system of (Liu and Zhang, 2017) only because of the definition of the **Open** action: our formulation pops from the stack the left corner constituent S_0 and pushes back on the stack the predicted category $\bullet N$ before S_0. That is, during parsing, the stack keeps the same internal structure as in the top down version (Dyer et al., 2016) but the tree traversal is different.

A derivation $(\mathbf{x}, \mathbf{y}) = a_1 \ldots a_m$ of an RNNG is a sequence of actions $a_1 \ldots a_m$. The weight of a derivation of length m is defined as

$$P(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{m} \log P(a_j | a_1 \ldots a_{j-1})$$

At each inference step the parser has to score the set of actions. The neural model seeks to assign a probability distribution $\mathbf{p}$ to the set A of actions given the parsing context:

$$\mathbf{p} = \textsc{Softmax}(\mathbf{Wh} + \mathbf{b})$$

The parsing context $\mathbf{h}$ is itself encoded by a stack-LSTM (Dyer et al., 2016):

$$\mathbf{h} = \textsc{Stack-Lstm}(\mathbf{e}_1 \ldots \mathbf{e}_n)$$

where each $\mathbf{e}_i$ is either a word $\mathbf{e}_i^W$, a nonterminal $\mathbf{e}_i^N$, or a tree embedding $\mathbf{e}_i^T$. Word representations are pushed on the stack-LSTM by the **generate** action and are the concatenation of a word embedding $\mathbf{w}$ and an embedding of the word's characters $c_1 \ldots c_n$:

$$\mathbf{e}_i^W = [\mathbf{w}; \textsc{Bi-Lstm}(c_1 \ldots c_n)]$$

Nonterminal $\mathbf{e}_i^N$ and tree embeddings $\mathbf{e}_i^T$ are computed using the same methods as in (Dyer et al., 2016).

As $\mathbf{p}$ encodes the conditional distribution $P(a_i | a_1 \ldots a_{i-1})$, the network is trained on treebank data of N sentences to maximize the log likelihood:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{m} \log P(a_j | a_1 \ldots a_{j-1})$$

2.2 Variable beam-size inference

Problem with naive beam search RNNG is initially formalized as a two stage parsing process (Dyer et al., 2016): first a discriminative model with lookahead is run on input sentences and then a generative model without lookahead is used to rerank the best hypotheses.

However direct generative parsing with RNNG is difficult because word generation transitions have significantly lower probabilities than structural transitions, such as open and close transitions. Thus, inside a beam at the same temporal step, these lexical transitions are likely to be pruned. This situation creates a bias towards parsing with additional spurious structure.

Some solutions already described in Stern et al. (2017) and Hale et al. (2018) manage to deal with this problem for direct generative parsing but this solution still requires to use rather large beams to yield accurate results, hence entailing significant computational overhead in time.

Among the alternative ways to solve the issue, the two stage architecture (Dyer et al., 2016) is not an option in our case because it relies on a discriminative model with lookahead that would break the strict incrementality property that we want to preserve for the cognitive perspective. Furthermore, in our scientific context, we have to study two inference problems: (1) the usual search for a best parse that is required for assessing the model accuracy with known metrics, and (2) for the purpose of fMRI modelling, we also have to compute the marginal probabilities required to compute the probabilities relevant for language modelling.

Proposed solution We therefore study another solution to this problem that is inspired by Sequential importance sampling methods (Doucet and Johansen, 2011; Buys and Blunsom, 2015) and that can be interpreted as a variable beam search strategy where the search method carefully avoids to compare apples and oranges within the same beam. Specifically, the method ensures that comparisons between derivations are made only if derivations have an identical number of generated lexical elements.

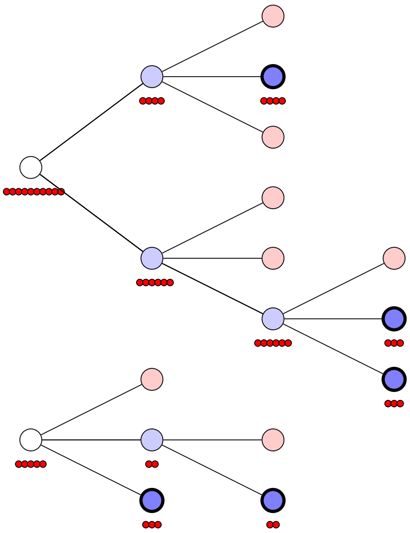

Figure 1: Example of a sampling step from derivations generating x_i (white nodes) to derivations generating x_{i+1} (circled blue nodes) with a budget of $K = 15$ particles. Inference may stop because of a lack of budget ($\pi(\mathbf{x}, \mathbf{y}) = 0$) as illustrated by red nodes. Derivations are never compared to each other during the sampling step, hence avoiding lexical biases that hamper the process of beam search.

For a sequence of words $\mathbf{x} = x_1 \ldots x_i \ldots x_n$, we note $\mathbf{x}_i$ the prefix $x_1 \ldots x_i$ and $\mathbf{y}_i = a_1 \ldots a_k$ a derivation whose last element generates x_i. We let $\mathcal{Y}(\mathbf{x}_i)$ denote the set of such derivations.

The search method essentially performs iteratively a move from word x_i to its successor x_{i+1} using the following procedure: given a distribution $P(\mathbf{x}_i, \mathbf{y}_i)$ over derivations at x_i, the algorithm samples the distribution $P(\mathbf{x}_{i+1}, \mathbf{y}_{i+1})$ with a swarm of particles. This procedure is decomposed in two steps:

Step 1: Sampling. Let ($\mathbf{x}$,$\mathbf{y}$) be a derivation together with its associated number of particles $\pi(\mathbf{x}, \mathbf{y})$, then each of its successors ($\mathbf{x}$,$\mathbf{y}a$) ($a \in A$) gets its particles assigned using the following recurrence[1]:

$$\pi(\mathbf{x}, \mathbf{y}a) = \lfloor \pi(\mathbf{x}, \mathbf{y}) P^*(a|\mathbf{x}, \mathbf{y}) \rceil \qquad (1)$$

This step involves expanding the set of derivations. Each derivation ($\mathbf{x}$,$\mathbf{y}$) is expanded iteratively until either the word x_{i+1} is generated or ($\mathbf{x}$,$\mathbf{y}$) has no more particle ($\pi(\mathbf{x}, \mathbf{y}) = 0$). This procedure is illustrated in Figure 1.

We have to emphasize, however, that contrary to $P(a|\mathbf{x}, \mathbf{y})$, the chosen importance distribution $P^*(a|\mathbf{x}, \mathbf{y})$ locally sums to one. It is indeed defined as:

$$P^*(a|\mathbf{x}, \mathbf{y}) = \frac{P(a|\mathbf{x}, \mathbf{y})}{\sum_{a' \in A^*} P(a'|\mathbf{x}, \mathbf{y})} \qquad (2)$$

where $A^* \subseteq A$ is a set of allowed actions given $(\mathbf{x}, \mathbf{y})$. Using the importance distribution prevents particles to get lost because of deficient probability distributions. Observe that $P(a|\mathbf{x}, \mathbf{y})$ does not sum to one because the generative model prevents a subset of actions to be undertaken at any time step. This is crucially the case for lexical actions when parsing: only the next word form given in the input sentence can be generated at any time step while there are lexical actions for generating every word in the vocabulary. This causes $P(a|\mathbf{x}, \mathbf{y})$ to be deficient.

A connection with importance sampling is established if we define the importance weight of a derivation as:

$$w(\mathbf{x}, \mathbf{y}) = \frac{P(\mathbf{x}, \mathbf{y})}{P^*(\mathbf{x}, \mathbf{y})}$$

where $P^*(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{m} P^*(a_j|a_1 \ldots a_m)$. We can then observe the relation between the three notions by developing:

$$\begin{aligned} P(\mathbf{x}, \mathbf{y}) &= \frac{P(\mathbf{x}, \mathbf{y})}{P^*(\mathbf{x}, \mathbf{y})} P^*(\mathbf{x}, \mathbf{y}) \\ &= w(\mathbf{x}, \mathbf{y}) P^*(\mathbf{x}, \mathbf{y}) \qquad (3) \end{aligned}$$

[1] We use $\lfloor x \rceil$ to denote the rounding of x to the nearest integer.

Step 2: Reweighting and **Filtering**. Observe, however, that the recurrence (1) spawns a search tree and that the mass of particles assigned to search branches gets increasingly scattered as the process progresses with further time steps. To counter this effect, we reweight the branches by synchronizing the search process on word generation events. In other words, each derivation $(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}) \in \tilde{\mathcal{Y}}(\mathbf{x}_{i+1})$ that successfully generated the word x_{i+1} is reweighted by normalizing its importance weight:

$$w'(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}) = \frac{w(\mathbf{x}_{i+1}, \mathbf{y}_{i+1})\pi(\mathbf{x}_{i+1}, \mathbf{y}_{i+1})}{\sum_{\mathbf{y}'} w(\mathbf{x}_{i+1}, \mathbf{y}')\pi(\mathbf{x}_{i+1}, \mathbf{y}')} \tag{4}$$

We subsequently reassign each derivation with a number of particles $\pi'(\mathbf{x}_{i+1}, \mathbf{y}_{i+1})$ proportional to its normalized weight. Let K be the global number of particles available, then we reallocate particles to each derivation as follows:

$$\pi(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}) = \lfloor K w'(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}) \rceil$$

Derivations without particles are filtered out at this stage ($\pi(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}) = 0$). Figure 2 summarizes the search algorithm with pseudo code. It can be interpreted as an adaptation of particle filtering (Doucet and Johansen, 2011) to our generative parsing task.

Marginal probabilities The proposed exploration method of the search space finds a direct application to computing marginal probabilities for sentential prefixes of the form $P(\mathbf{x}_i)$. For our purposes, one key interest of the generative parsing model lies in its capacity to compute transition probabilities $P(x_i|x_1 \dots x_{i-1})$ informed by the syntactic structures of the sentence. This enables the computation of metrics that have been shown relevant for psycho-linguistic analysis such as surprisal and entropy, among others.

We briefly summarize in the following how this is achieved through the computation of marginal probabilities. We can define the probability of a prefix as:

$$P(\mathbf{x}_i) = \sum_{\mathbf{y}_i \in \mathcal{Y}(\mathbf{x}_i)} P(\mathbf{x}_i, \mathbf{y}_i) \tag{5}$$

Thus computing transition probabilities can be simply performed as:

$$P(x_i|x_1 \dots x_{i-1}) = \frac{P(\mathbf{x}_i)}{P(\mathbf{x}_{i-1})} \tag{6}$$

```
function VariableBeamSearch($x_1 \dots x_n$,$K$)
  $\tilde{\mathcal{Y}}(\mathbf{x}_0) \leftarrow (\epsilon, \epsilon)$
  for $i \in 0 \dots n-1$ do
    agenda $\leftarrow \tilde{\mathcal{Y}}(\mathbf{x}_i)$                          ▷ Sampling step
    $\tilde{\mathcal{Y}}(\mathbf{x}_{i+1}) \leftarrow \emptyset$
    while agenda $\neq \emptyset$ do
      $(\mathbf{x}, \mathbf{y}) \leftarrow$ Pop(agenda)
      for $a \in A$ do
        $\pi(\mathbf{x}, \mathbf{y}a) \leftarrow \lfloor \pi(\mathbf{x}, \mathbf{y}) P^*(a|\mathbf{x}, \mathbf{y}) \rceil$
        if $\pi(\mathbf{x}, \mathbf{y}a) > 0$ then
          if $a$ generates $x_{i+1}$ then
            $\tilde{\mathcal{Y}}(\mathbf{x}_{i+1}) \leftarrow \tilde{\mathcal{Y}}(\mathbf{x}_{i+1}) \cup (\mathbf{x}a, \mathbf{y})$
          else
            agenda $\leftarrow$ agenda $\cup (\mathbf{x}, \mathbf{y}a)$
          end if
        end if
      end for
    end while
    for $(\mathbf{x}, \mathbf{y}) \in \tilde{\mathcal{Y}}(\mathbf{x}_{i+1})$ do                ▷ Reweighting
      $w(\mathbf{x}, \mathbf{y}) \leftarrow \frac{w(\mathbf{x}, \mathbf{y})\pi(\mathbf{x}_{i+1}, \mathbf{y}_{i+1})}{\sum_{\mathbf{y}' \in \tilde{\mathcal{Y}}(\mathbf{x}_{i+1})} w(\mathbf{x}, \mathbf{y}')\pi(\mathbf{x}_{i+1}, \mathbf{y}')}$
    end for
    for (x,y) $\in \tilde{\mathcal{Y}}(\mathbf{x}_{i+1})$ do                   ▷ Filtering
      $\pi(\mathbf{x}, \mathbf{y}) \leftarrow \lfloor K w(\mathbf{x}, \mathbf{y}) \rceil$
    end for
    $\tilde{\mathcal{Y}}(\mathbf{x}_{i+1}) \leftarrow \{(\mathbf{x}, \mathbf{y}) | \pi(\mathbf{x}, \mathbf{y}) > 0, (\mathbf{x}, \mathbf{y}) \in \tilde{\mathcal{Y}}(\mathbf{x}_{i+1})\}$
  end for
  return $\tilde{\mathcal{Y}}(\mathbf{x}_n)$
end function
```

Figure 2: Variable beam search pseudo-code

The naive computation of marginal probabilities stated in (5) requires to process a set $\mathcal{Y}(\mathbf{x}_i)$ growing exponentially in size with the length of the input sequence. Such a computation can however be approximated by Monte-Carlo simulation:

$$P(\mathbf{x}_i) \approx \lim_{K \to \infty} \frac{1}{K} \sum_{j=1}^{K} w(\mathbf{x}_i, \mathbf{y}_i^j) \tag{7}$$

This involves sampling derivations $\mathbf{y}_i^1 \dots \mathbf{y}_i^K$ from $P^*(\mathbf{x}, \mathbf{y})$, and is justified by observing that the marginalisation of equation (3) actually states an expectation:

$$P(\mathbf{x}_i) = \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{Y}(\mathbf{x}_i)} w(\mathbf{x}, \mathbf{y}) P^*(\mathbf{x}, \mathbf{y}) \tag{8}$$

which is precisely the one computed by (7). As our exploration method does not directly samples K individual derivations but clusters the K particles on a smaller number of derivations with the recurrence (1). We reformulate (8) not directly as (7) but rather as:

$$\begin{aligned}P(\mathbf{x}_i) &= \sum_{(\mathbf{x},\mathbf{y})\in\tilde{\mathcal{Y}}(\mathbf{x}_i)} w(\mathbf{x},\mathbf{y})P^*(\mathbf{x},\mathbf{y})\\ &= \sum_{(\mathbf{x},\mathbf{y})\in\tilde{\mathcal{Y}}(\mathbf{x}_i)} P(\mathbf{x},\mathbf{y})\frac{P^*(\mathbf{x},\mathbf{y})}{P^*(\mathbf{x},\mathbf{y})}\\ &= \sum_{(\mathbf{x},\mathbf{y})\in\tilde{\mathcal{Y}}(\mathbf{x}_i)} P(\mathbf{x},\mathbf{y})\end{aligned}$$

That is we approximate the sum on the full set of parses $\mathcal{Y}(\mathbf{x}_i)$ by the sum of the set of parses found in the beam $\tilde{\mathcal{Y}}(\mathbf{x}_i)$ and sampled with the importance distribution $P^*(\mathbf{x},\mathbf{y})$.

Parsing problem Actual parsing is performed by collecting all successful derivations $\mathrm{succ}(\mathbf{x})$ found during the search process and by returning the best scoring derivation:

$$(\mathbf{x},\hat{\mathbf{y}}) = \operatorname*{argmax}_{(\mathbf{x},\mathbf{y})\in\mathrm{succ}(\mathbf{x})} P(\mathbf{x},\mathbf{y})$$

2.3 Measures extracted from the parser

We now define several measures extracted from the parsing process that might be relevant for fMRI modelling.

The first is the **beam size**, that is the size of the beam at word x_i and it is directly measured as $B = |\tilde{\mathcal{Y}}(\mathbf{x}_i)|$

Next we define the **beam successful activity**. Let the set of transitions of a derivation be: $T(\mathbf{y}) = T(a_1 \dots a_m) = \{(a_1,a_2)\dots(a_{m-1},a_m)\}$ and the set of transitions $\mathbf{S}(\mathbf{x}_i) = \bigcup_{\mathbf{y}\in\tilde{\mathcal{Y}}(\mathbf{x}_i)} T(\mathbf{y})$ be the set of transitions leading successfully to x_i.

Then, the beam successful activity at x_i is defined as the size of the set of transitions within x_i and x_{i-1}, that is :

$$s(x_i) = |\,\mathbf{S}(\mathbf{x}_i) - [\mathbf{S}(\mathbf{x}_i)\cap\mathbf{S}(\mathbf{x}_{i-1})]\,|$$

Intuitively, the beam successful activity metric measures the amount of nodes depicted in blue in Figure 1.

A dead-end is a derivation $\mathbf{y}$ such that $\pi(\mathbf{y}) > 0$ and $\pi(\mathbf{y}a) = 0 \quad (\forall a \in A)$. Let $\mathcal{D}(\mathbf{x}_i)$ be the set of such dead-ends between x_{i-1} and x_i and the set of transitions $\mathbf{D} = \bigcup_{\mathbf{y}\in\tilde{\mathcal{D}}(\mathbf{x}_i)} T(\mathbf{y})$, then we define the **beam unsuccessful activity** as:

$$u(x_i) = |\,\mathbf{D}(\mathbf{x}_i) - [\mathbf{D}(\mathbf{x}_i)\cap\mathbf{S}(\mathbf{x}_{i-1})]\,|$$

Intuitively, the beam unsuccessful activity metric measures the amount of nodes depicted in red in Figure 1. Finally, let $\mathbf{A}(\mathbf{x}_i) = \mathbf{S}(\mathbf{x}_i)\cup\mathbf{D}(\mathbf{x}_i)$, then we define an overall **beam activity** at x_i to be:

$$o(x_i) = |\,\mathbf{A}(\mathbf{x}_i) - [\mathbf{A}(\mathbf{x}_i)\cap\mathbf{S}(\mathbf{x}_{i-1})]\,|$$

As our model is generative, information theoretic metrics, such as surprisal and entropy, that are crucially dependant of the computation of lexical transition probabilities of the form $P(x_i|x_1\dots x_{i-1})$ can be computed too. **Surprisal** is a classic complexity measure (Hale, 2016) defined as:

$$\textbf{surprisal}(x_i) = -\log_2 P(x_i|x_1\dots x_{i-1})$$

and that is computed with equation (6).

For **entropy**, let $H[\mathcal{Y}(\mathbf{x}_i)]$ be the entropy of the set of prefix derivations at word i that is defined as:

$$H[\mathcal{Y}(\mathbf{x}_i)] = -\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x}_i)} \frac{P(\mathbf{x}_i,\mathbf{y})}{P(\mathbf{x}_i)}\log_2\left(\frac{P(\mathbf{x}_i,\mathbf{y})}{P(\mathbf{x}_i)}\right)$$

However in the context of a beam with variable size, these entropy measures cannot be directly compared at different time steps $i, i', i'' \dots$ because the size of the set $\mathcal{Y}(\mathbf{x}_i)$ is variable accross time steps. Therefore we use a **normalized entropy** whose values range in the interval $[0,1]$. As the maximum of the entropy measured on a set of n elements is $\log(n)$, we use the normalized entropy measure:

$$H_n[\mathcal{Y}(\mathbf{x}_i)] = \frac{H[\mathcal{Y}(\mathbf{x}_i)]}{\log_2(B)}$$

To conclude this section, we introduced a search method for generative neural parsing that aims to overcome the problems of lexical biases (Stern et al., 2017) by drawing inspiration from particle filtering methods. In section 3 we motivate and study some properties of the method from a computational perspective and we explore in section 5 the potential of the overall beam activity predictors for modelling and analyzing fMRI data (Table 3).

3 Computational experiments

We report a few experiments designed to better understand the computational properties of the search method described so far. All the experiments are performed on the Penn-Treebank (Marcus et al., 1994), using sections 2-21 as training, section 22 as development and section 23 as test.

We preprocess the data for numbers replaced by a unique `num` token and we replace word occurrences with frequency one by the token `unk`.

Our development experiments compare our variable-size beam method with more traditional beam search. As standard naive beam search is not appropriate to generative parsing, we directly compare with the word-synchronous beam search method of Stern et al. (2017), applied to RNNG, with the so-called fast track extension. Our implementation of word synchronous beam search follows closely that of Hale et al. (2018). We select the settings found in the literature: for a beam of size $B = k$, a lexical beam of size $k/10$ and a fast-track of size $k/100$.

MODEL	F-SCORE	PPL	BEAM ACTIVITY
K=1000-base	86.29	107.89	153.17
K=1000	90.75	86.00	66.3
K=10000	91.12	84.25	153.6
K=50000	**91.26**	83.50	380.3
B=100	87.21	95.97	416.9
B=400	90.50	**82.39**	1775.1

Table 1: Development scores.

By observing the development measures in Table 1, the first thing to note is the lack of accurracy of the vanilla application of our *base* method, as can be seen on the first line of the Table, when parsing with $K = 1000$ particles for searching. This *base* model uses straightforwardly the reweighting step stated in (4). We observed however that the low accurracy of this model is a consequence of a rounding issue when using discretized particle counts. To avoid this problem we used the alternative reweighting scheme:

$$w'(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}) = \frac{P(\mathbf{x}_{i+1}, \mathbf{y}_{i+1})}{\sum_{\mathbf{y}' \in \tilde{\mathcal{Y}}(\mathbf{x}_{i+1})} P(\mathbf{x}_{i+1}, \mathbf{y}')}$$

that avoids this rounding problem and all the other results reported in Table 1 use instead this alternative scheme.

From the development set we can also point out that our beam baselines are almost identical to those of Hale et al. (2018) on beams $B = 100$ and $B = 400$. Note that the authors managed to reach an F-score of 91.2 on the development with $B = 2000$. By comparison, our particle method tends to produce high F-scores even with limited K. Although the perplexity of the beam method remains generally lower, as shown in Table 1.

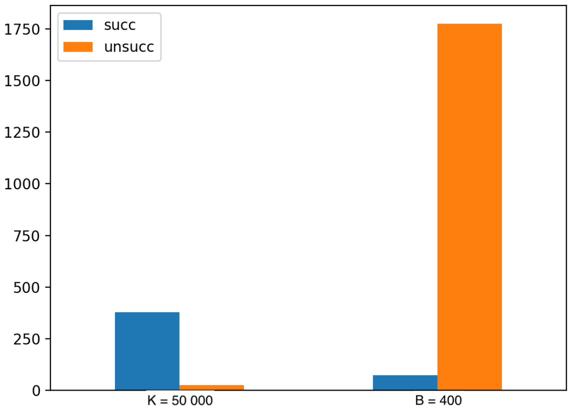

Figure 3: Successful and unsuccessful beam activities.

The main difference comes from computation time that we measure with an averaged beam activity metric defined earlier on: as we can see our most expensive model $K = 50000$ requires to perform less computations than the smallest beam tested. This striking difference can be further explained by comparing the successful and unsuccessful beam activities.

Figure 3 plots those activities for models $B = 400$ and $K = 50000$. The more traditional beam search spends most of its time performing unsuccessful computations while our method spends most of its time performing successful computations. All these observations have practical consequences: the beam method takes hours to complete the parsing of the Wall Street Journal (wsj, see Table 2) development while the particle search takes minutes on a standard workstation.

MODEL	F-SCORE	PPL (wsj)	PPL (prince)
K=50000	**91.02**	94.35	154.93
B=400	90.15	**93.02**	139.76
IKN5	-	155.02	309.54
LSTM-LM	-	141.28	204.06
(Dyer et al. 2016)	93.3	105.2*	unknown
(Fried et al. 2017a)	92.56	unknown	unknown
(Kitaev et al. 2018)	**93.55**	-	-

∗A different preprocessing is used. The comparison remains indicative

Table 2: Test results.

For the test, we observe again from Table 2 that our search method ($K = 50000$) gets better F-score and worse perplexity than the beam method.

The most important observation comes from the comparison between the language model perplexities and parsing perplexities. On a small data set such as the Penn Treebank, the parser has a much better perplexity than traditional 5-grams Interpolated Kneser Ney (IKN5) or vanilla LSTM-LM with a comparable configuration to that of the parser. It is quite clear that state of the art language models outperform the parsing language model by a large margin, at the cost of much larger training sets (Radford et al., 2019). The perplexities reported as *Prince* in Table 2 are the perplexities resulting from processing the *Little Prince* corpus used later in the paper for analyzing neuro-imaging data. The table allows to measure the amount of drop in perplexity that can be explained by corpus domain change.

We also compare with parsers without strict incrementality restrictions: Dyer et al. (2016) is an RNNG as a reranker with a discriminative first step (that is with lookahead). Stern et al. (2017) are the results reported with a beam of size 2000 using the generative model described by Choe and Charniak (2016) with a substantially different preprocessing of the data than the one used with RNNG (Hale et al., 2018). Finally, Kitaev and Klein (2018) is the current state of the art single parser with a discriminative model.

4 Neuro-imaging and Parsing Models

A number of studies observed a positive correlation between behavioural and psycho-linguistic data like reading times and entropy measures derived from probabilistic parsers' computations (Levy, 2008). Eye-movements (Demberg and Keller, 2008) and Event-Related Potentials (ERPs, Frank et al. (2015) were recorded during reading performance or sentence completion tasks have allegedly contributed in giving a cognitive dimension to the Surprisal Theory (Levy, 2008).

Parsers Actions and Neuro-imaging Further empirical proofs of the link between information-theoretic measures and cognitive processes showed a correlation between the *output* or internal state of syntactic parsers and cerebral activity. In a seminal work based on a 30 minutes recording from the English text *Alice in wonderland*, Brennan and colleagues (2012; 2016) modeled brain activity with a simple word-by-word measure of Node counts, featuring an estimate of the amount syntactic structure analyzed so far, as the number of phrases closed at each word. This computationally assessed phrase-structure building process showed to involve Inferior Frontal Gyrus and Anterior Temporal regions, which was found consistent with earlier studies on sentence structure building (Pallier et al., 2011; Snijders et al., 2009).

Parsing strategy in Neuro-imaging Other fMRI studies started to use another computationally derived complexity metric to quantify structure-building and processing effort in the brain. The number of rules applied during the parser's computational procedure, between each word of a sentence, was taken to define an incremental index of computational syntactic "work"(e.g. Bhattasali et al. 2019).

A recent intracranial recording work correlated brain activity with computational metrics from alternative grammar types (e.g. context-free, Minimalist) and different parsing strategies (e.g. left-corner, top-down, bottom-up) (Nelson et al., 2017). Phrase-structure appeared, like in more traditional approaches (Friederici and Gierhan, 2013; Snijders et al., 2009; Hickok and Poeppel, 2007), as a major determinant of the dynamic profile of brain activity in language areas. Superior temporal and inferior frontal areas provided a good fit of the brain-activity data for bottom-up and left-corner parsing strategies, in left superior temporal, inferior frontal, dorsal and midline frontal sites.

Based on this set of empirical evidence, showing that syntactic structure-building implicates frontal regions, such as the Pars Triangularis and Pars Opercularis of the Inferior Frontal Gyrus (IFG-Broca) and anterior temporal areas, neuro-imaging signals can be used nowadays to adjudicate between competing parsing mechanisms. The present paper, goes a step further, using the beam internal state in two different Search methods in a generative RNNG architecture, to shed light on the potential benefits of beam-size variation for the cognitive modelling of sentence processing.

5 fMRI: Method and Results

The analyzed text was the transcribed version of the English audio-book of Antoine de Saint-Exupéry *The Little Prince*, translated by David Wilkinson and read by Nadine Eckert-Boulet. This text comprises 19,171 tokens and 15,388 words, grouped in 1388 sentences. The imaging data-set analysed in this paper comprises 50

Predictors	Description
Overall Beam activity K50000 or B400	word-by-word measure of beam activity (§2.3)
Word rate	tagging spoken words on the fMRI signal
Word frequency	word-by-word log-frequency in movie subtitles
RMS amplitude	an acoustic correlate of volume every 10ms

Table 3: Predictors used in the fMRI Analysis. RMS (Root Mean Square) encodes the amplitude of spoken narration, it reflects the intensity.

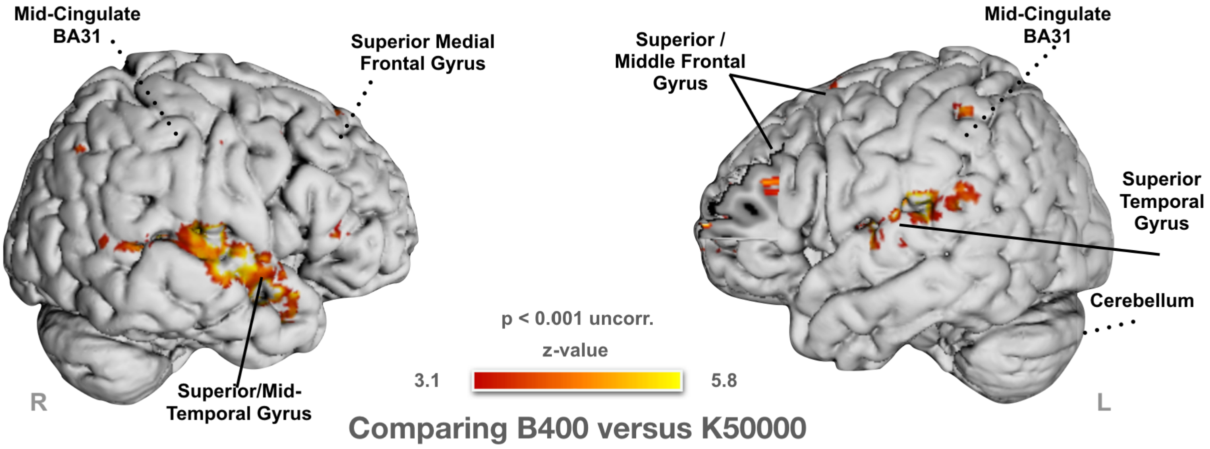
Mid-Cingulate BA31
Superior Medial Frontal Gyrus
Superior/Mid-Temporal Gyrus
R
p < 0.001 uncorr.
z-value
3.1
5.8
Comparing B400 versus K50000
Superior / Middle Frontal Gyrus
Mid-Cingulate BA31
Superior Temporal Gyrus
Cerebellum
L

Figure 4: Brain z-maps showing the significant clusters (K>25) for the model comparison between Overall Beam Activity calculated in the fix-beam B400 versus variable-beam K50000 RNNG parser, p < .001 uncorr.

right-handed native English participants who listened to the entire audio-book during 1h38 minutes. Imaging was performed in a 3 Tesla MRI scanner (Discovery MR750 GE, Milwaukee, WI) with a 32-channel head coil at the Cornell MRI Facility as part of a naturalistic listening neuro-imaging project. Details of preprocessing are presented in the Supplementary Materials.

5.1 Statistical Analysis

r^2 Model comparison - General Linear Model The research questions presented above in section 2.2 motivates a statistical analysis that performs a comparison where fMRI signal is modeled (GLM) by B400 fix beam parser versus the K50000 parsing system adopting a variable beam.

Single-subject statistics At the single subject level, the observed time-course of the brain hemodynamic response (BOLD - Blood Oxygenation Level Dependent) in each voxel was modeled by the Overall Beam Activity measure (cf. 2.3) calculated for B400 and K50000 parsers, and time-locked at the offset of each word in the audio-book. Model comparisons using cross-validated coefficient of determination (r2) maps was carried out in order to evaluate the goodness of fit of the two beam search methods with BOLD signal. The predictors shown in Table 3 were convolved using SPM's (Friston et al., 2007) canonical HRF (Hemodynamic Response Function). The two neuro-imaging models (i.e. K50000 and B400) also included in the GLM three control variables: *Root Mean Square intensity* (RMS) an indicator of intensity at every 10 ms of the audio; *word rate* as a stick function marking the offset of each spoken word; *frequency* of the individual words in movie subtitles from (Brysbaert and New, 2009).

Such predictors are here to ensure that conclusions about parsing difficulty would be specific to the processes they instantiate, as opposed to more general aspects of speech perception. Specifically, lexical frequency was added as a covariate of non-interest, to statistically factor out effects of general word frequency (generally used in psycholinguistics), that may correlate with other types of expectations. For every subject, we compute how much the inclusion of each variable of interest (i.e. B400 and K50000) increases the cross-validated r^2, from a baseline model that does not include them, see Table 3). Hence, the r^2 scores represent the variance explained in each voxel by the variable instantiating the two beam-search methods.

Regions for K50000 vs B400	Cluster size	MNI Coordinates			p-value	z-score
	(in voxels)	x	y	z	(uncorrected)	(peak)
R Superior / Mid-Temporal Gyrus	2307	54	-8	-2	0.000	5.79
L Superior Temporal Gyrus / Insula	1442	-52	-12	2	0.000	5.73
R Superior Medial Frontal Gyrus (BA9/10)	114	4	56	24	0.000	4.89
R Mid-Cingulate Gyrus (BA31)	95	4	-42	40	0.000	4.30
L Middle Frontal Gyrus	58	-40	36	22	0.000	4.13
L Cerebellum - Crus I/II	34	-22	-74	-36	0.000	4.02
L Superior Frontal Gyrus (10)	34	-28	58	2	0.001	3.64
L Mid-Cingulate Gyrus (BA31)	95	-6	-28	44	0.001	3.66

Table 4: Clusters showing a significant better fit for B400 fix beam-size search method in the RNNG parser, p< 0.001 (z-score > 3.1) uncorrected at k> 25 cluster threshold.

Group-level statistics To compare the impact of beam-search on fMRI signal explanation (i.e. r^2 increase of each variable) we performed a paired t-test on each individual r^2 map, and obtained Figure 4 showing where one variable explains significantly better the signal than the other (cf. Tab. 4).

5.2 Results - Fit with fMRI signal

We performed an r^2 comparison to test which beam search method provided the better fit to the fMRI data recorded during *The Little Prince*. The two different search methods were tested (K50000 variable beam and B400 fix beam), and B400 was shown to be the best fitting the BOLD signal of these models. Figure 4 (clusters coordinates and statistics, cf. Table 4), shows the significance (z-scores, $p < 0.001$ uncorrected) of the difference in r^2 scores with a cluster threshold of 25 voxels. Considering the cluster with a larger extent, we might also have captured some acoustic parameter that was not included in the GLM model, although further analyses showed that adding a prosodic predictor does not have a major impact.

Of the two parsers' search methods, the fix beam one had a significant predictive value in well-known language areas, namely temporal areas and sub-parts core frontal regions. Our particle filtering based method is used to maintain the beam and limit its size, while experiments show that the proposed method achieves higher parsing accuracy than a fixed-size beam search baseline, the fixed-size beam search is shown here to be a better model than the proposed method in predicting brain activities. This result could be related to its tendency to keep ambiguities longer along the sentence, and thus model the computational load of a larger spectrum of hypotheses without pruning immediately the hypothesis space (cf. Fig. 3).

6 Discussion & Conclusion

The paper investigates the relevance of a sequential Monte-Carlo inspired search method for generative parsing. The proposed search method is in principle very general although we tested its relevance to overcome lexical biases inherent to direct generative parsing for RNNG. We found out that the method's main benefit is an increased processing efficiency. The measures instantiating variable-beam and fixed-beam search were used to quantify the amount of structure-building ”work” the parser performs in the course of word-by-word processing of a given sentence. Our current neuro-imaging results suggest that the fixed-size beam activity is a better predictor of brain activity than the variable-size beam activity. The original methodology presented in this paper paves the way for further computational work in quantifying parsing complexity and thus fine-grained modelling of human sentence processing.

Acknowledgments

We are indebted to Wenming Luh, Nathan Spreng and John Hale for collaborating and providing fMRI data. We thank Shohini Bhattasali and Jixing Wang for assistance in data collection and preprocessing. We also thank the anonymous reviewers for their detailed comments on the original manuscript. This work is supported by the French National Research Agency (ANR) under grant ANR-14-CERA-0001, we gratefully acknowledge the National Science Foundation under Grant No.1607441 for data collection.

References

Shohini Bhattasali, Murielle Fabre, Wen-Ming Luh, Hazem Al Saied, Mathieu Constant, Christophe Pallier, Jonathan R. Brennan, R. Nathan Spreng, and John Hale. 2019. Localising memory retrieval and syntactic composition: an fmri study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, 34(4):491–510.

Jonathan Brennan. 2016. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.

Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J Heeger, and Liina Pylkkänen. 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, 120(2):163–173.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–990.

Jan Buys and Phil Blunsom. 2015. Generative incremental dependency parsing with neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 863–869.

Do Kook Choe and Eugene Charniak. 2016. Parsing as language modeling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2331–2336.

Vera Demberg and Frank Keller. 2008. Data from eyetracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193 – 210.

Arnaud Doucet and Adam M Johansen. 2011. *A tutorial on particle filtering and smoothing: Fifteen years later*, d. crisan and b. rozovsk, eds. edition, volume 12 (3) of *Handbook on Nonlinear Filtering*, pages 656–704. Oxford Press.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent neural network grammars.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1 – 11.

Angela D Friederici and Sarah ME Gierhan. 2013. The language network. *Current Opinion in Neurobiology*, 23(2):250–254.

K.J. Friston, J. Ashburner, S.J. Kiebel, T.E. Nichols, and W.D. Penny, editors. 2007. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press.

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

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2726–2735.

Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2675–2685.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A Smith. 2017. What do recurrent neural network grammars learn about syntax?

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):11261177.

Jiangming Liu and Yue Zhang. 2017. In-order transition-based constituent parsing. volume 5, pages 413–424.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 114–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew J. Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S. Cash, Lionel Naccache, John T. Hale, Christophe Pallier, and Stanislas Dehaene. 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Unpublished manuscript.

Tineke M Snijders, Theo Vosse, Gerard Kempen, Jos JA Van Berkum, Karl Magnus Petersson, and Peter Hagoort. 2009. Retrieval and unification of syntactic structure in sentence comprehension: An fMRI study using word-category ambiguity. *Cerebral Cortex*, 19(7):1493–1503.

Mitchell Stern, Daniel Fried, and Dan Klein. 2017. Effective inference for generative neural parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1695–1700.