Grammar Acquisition using Human-like Memory Bounds

William Schuler
Dept. of Linguistics, The Ohio State University

April 26, 2016

People have bounded memory for sentence processing

A history of results show recall declines if constituents center-embed (Miller & Isard, 1964, inter alia):

(1) The jeweler made [the ring that won [the prize that was given at the fair]].
(sentence recalled correctly)

(2) [The prize that [the ring that [the jeweler] made] won] was given at the fair.
(sentence recalled less accurately)

People have bounded memory for sentence processing

A history of results show recall declines if constituents center-embed (Miller & Isard, 1964, inter alia):

(1) The jeweler made [the ring that won [the prize that was given at the fair]].
(sentence recalled correctly)

(2) [The prize that [the ring that [the jeweler] made] won] was given at the fair.
(sentence recalled less accurately)

There are more common examples, but still difficult:

(3) If [either [both [the power is on] and the door is closed] or the power is off] then the bell will stop.

People have bounded memory for sentence processing

Corpus results show human language is highly depth-constrained (Karlsson, 2007; Schuler et al., 2010):

Switchboard (transcribed spontaneous speech):

memory capacity (no punct)	sentences	coverage
no stack memory	26,201	28.38%
1 stack element	53,740	58.21%
2 stack elements	85,068	92.14%
3 stack elements	91,890	99.53%
4 stack elements	92,315	99.99%
5 stack elements	92,328	100.00%
TOTAL	92,328	100.00%

People have bounded memory for sentence processing

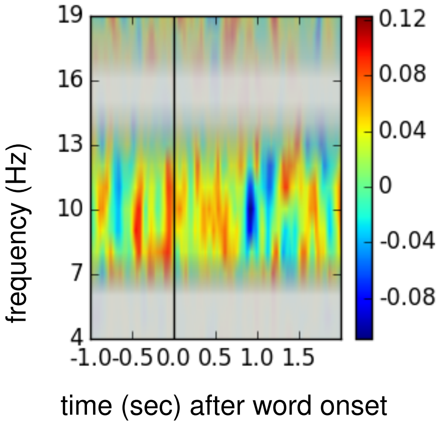
Corpus results show human language is highly depth-constrained (Karlsson, 2007; Schuler et al., 2010):

Wall Street Journal (newspaper text):

memory capacity (no punct)	sentences	coverage
no stack memory	127	0.32%
1 stack element	3,550	8.90%
2 stack elements	25,948	65.06%
3 stack elements	38,948	97.66%
4 stack elements	39,866	99.96%
5 stack elements	39,883	100.00%
TOTAL	39,883	100.00%

People have bounded memory for sentence processing

EEG show increased theta, gamma coherence during center embedding (left anterior-posterior), consistent with storage of incomplete signs (Weiss et al., 2005; van Schijndel et al., 2015):

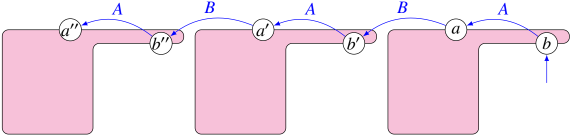


Left-corner parsers model human-like memory bounds

(Johnson-Laird, 1983; Abney & Johnson, 1991; Gibson, 1991; Resnik, 1992; Stabler, 1994; Lewis & Vasishth, 2005)

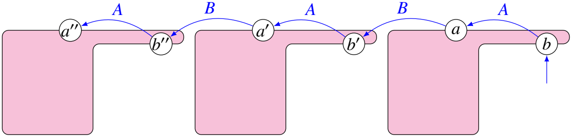
Left-corner parsers model human-like memory bounds

(Johnson-Laird, 1983; Abney & Johnson, 1991; Gibson, 1991; Resnik, 1992; Stabler, 1994; Lewis & Vasishth, 2005)

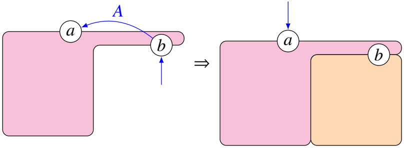


Left-corner parsers model human-like memory bounds

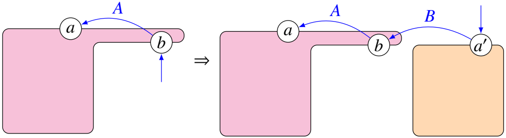
(Johnson-Laird, 1983; Abney & Johnson, 1991; Gibson, 1991; Resnik, 1992; Stabler, 1994; Lewis & Vasishth, 2005)



No-fork option:

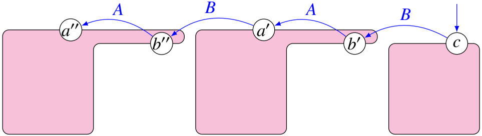


Fork option:



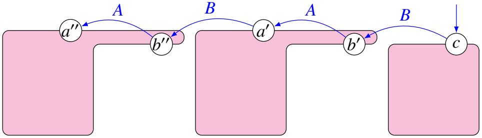
Left-corner parsers model human-like memory bounds

(Johnson-Laird, 1983; Abney & Johnson, 1991; Gibson, 1991; Resnik, 1992; Stabler, 1994; Lewis & Vasishth, 2005)

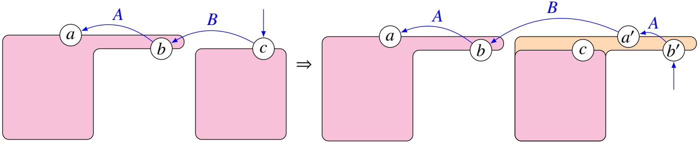


Left-corner parsers model human-like memory bounds

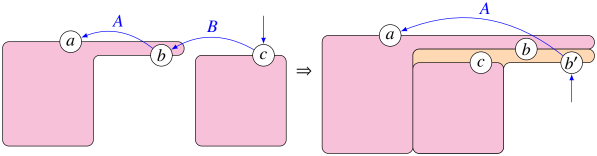
(Johnson-Laird, 1983; Abney & Johnson, 1991; Gibson, 1991; Resnik, 1992; Stabler, 1994; Lewis & Vasishth, 2005)



No-join option:

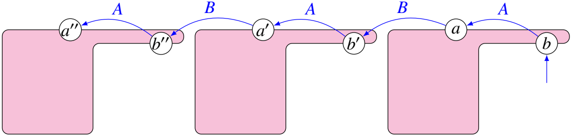


Join option:



Left-corner parsers model human-like memory bounds

(Johnson-Laird, 1983; Abney & Johnson, 1991; Gibson, 1991; Resnik, 1992; Stabler, 1994; Lewis & Vasishth, 2005)

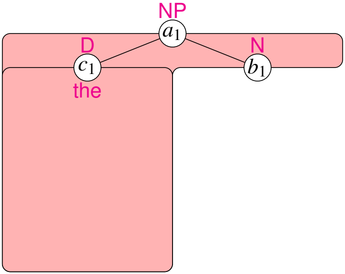


Left-corner parsers model human-like memory bounds

For example: start with zero incomplete signs

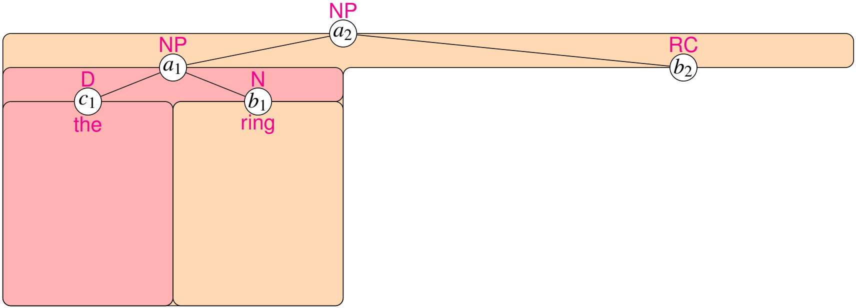
Left-corner parsers model human-like memory bounds

For example: fork, no join (now one incomplete sign: NP/N)



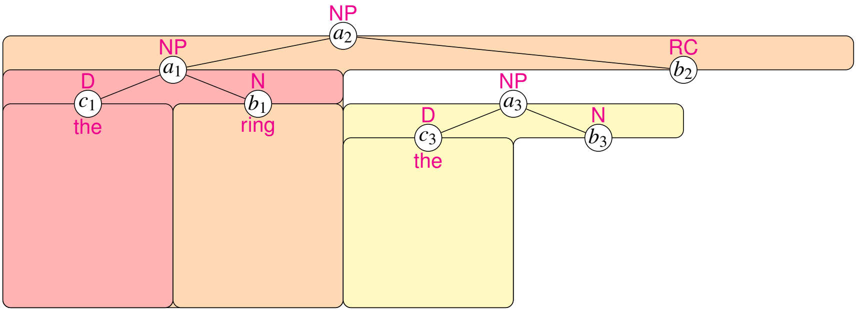
Left-corner parsers model human-like memory bounds

For example: no fork, no join (still <u>one</u> incomplete sign: NP/RC)



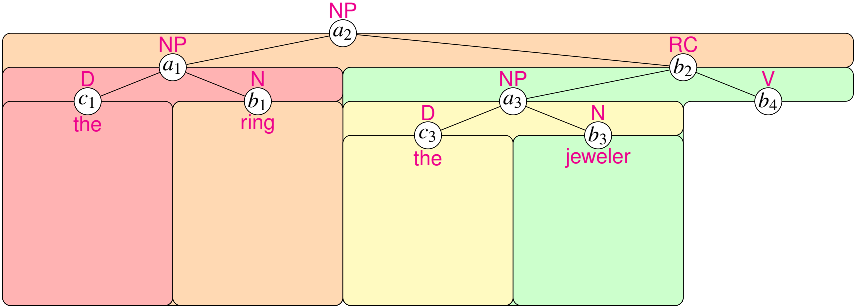
Left-corner parsers model human-like memory bounds

For example: fork, no join (now two incomplete signs: NP/RC, NP/N)



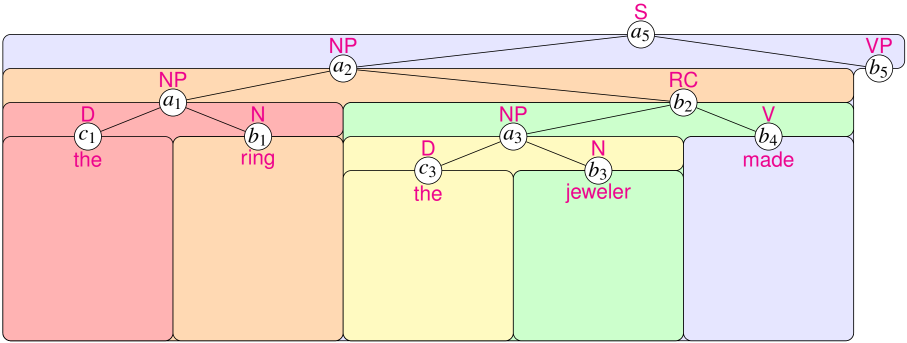
Left-corner parsers model human-like memory bounds

For example: no fork, join (now one incomplete sign: NP/V)



Left-corner parsers model human-like memory bounds

For example: no fork, no join (still one incomplete sign: S/VP)



Left-corner parsing is as accurate as bottom-up parsing

van Schijndel et al. (2013):

System	Precision	Recall	F-score
Roark 2001 (CNF)	86.6	86.5	86.5
Left-corner (CNF, beam width 500)	86.6	87.3	87.0
Left-corner (CNF, beam width 2000)	87.8	87.8	87.8
Left-corner (CNF, beam width 5000)	87.8	87.8	87.8
Petrov Klein (CNF)	88.1	87.8	88.0
Petrov Klein (not CNF)	88.3	88.6	88.5

Left-corner parsing is efficient with HHMMs

$\mathsf{P}(\vec{s}_t\, w_t \mid \vec{s}_{1..t-1}\, w_{1..t-1}) = \mathsf{P}(\vec{s}_t\, w_t \mid \vec{s}_{t-1})$ Markov assump.

Left-corner parsing is efficient with HHMMs

$$\mathsf{P}(\vec{s}_t\, w_t \mid \vec{s}_{1..t-1}\, w_{1..t-1}) = \mathsf{P}(\vec{s}_t\, w_t \mid \vec{s}_{t-1})$$ Markov assump.

$$= \mathsf{P}(\vec{s}_t \mid \vec{s}_{t-1}) \cdot \mathsf{P}(w_t \mid \vec{s}_t)$$ transition & emit

Left-corner parsing is efficient with HHMMs

$$
\begin{aligned}
\mathsf{P}(\vec{s}_t\, w_t \mid \vec{s}_{1..t-1}\, w_{1..t-1}) &= \mathsf{P}(\vec{s}_t\, w_t \mid \vec{s}_{t-1}) && \text{Markov assump.} \\
&= \mathsf{P}(\vec{s}_t \mid \vec{s}_{t-1}) \cdot \mathsf{P}(w_t \mid \vec{s}_t) && \text{transition \& emit} \\
&= \mathsf{P}(f_t\, j_t\, \vec{a}_t\, \vec{b}_t\, p_t \mid \vec{s}_{t-1}) \cdot \mathsf{P}(w_t \mid p_t) && \text{left-corner trans.}
\end{aligned}
$$

Left-corner parsing is efficient with HHMMs

$$\begin{aligned}
\mathsf{P}(\vec{s}_t\, w_t \mid \vec{s}_{1..t-1}\, w_{1..t-1}) &= \mathsf{P}(\vec{s}_t\, w_t \mid \vec{s}_{t-1}) && \text{Markov assump.}\\
&= \mathsf{P}(\vec{s}_t \mid \vec{s}_{t-1}) \cdot \mathsf{P}(w_t \mid \vec{s}_t) && \text{transition \& emit}\\
&= \mathsf{P}(f_t\, j_t\, \vec{a}_t\, \vec{b}_t\, p_t \mid \vec{s}_{t-1}) \cdot \mathsf{P}(w_t \mid p_t) && \text{left-corner trans.}
\end{aligned}$$

Unconstrained, this would cost $\mathcal{O}(K^{2\times D\times 2})$,
(K is the number of categories, D is the maximum depth).

Left-corner parsing is efficient with HHMMs

$$
\begin{aligned}
\mathsf{P}(\vec{s}_t\, w_t \mid \vec{s}_{1..t-1}\, w_{1..t-1}) &= \mathsf{P}(\vec{s}_t\, w_t \mid \vec{s}_{t-1}) && \text{Markov assump.}\\
&= \mathsf{P}(\vec{s}_t \mid \vec{s}_{t-1}) \cdot \mathsf{P}(w_t \mid \vec{s}_t) && \text{transition \& emit}\\
&= \mathsf{P}(f_t\, j_t\, \vec{a}_t\, \vec{b}_t\, p_t \mid \vec{s}_{t-1}) \cdot \mathsf{P}(w_t \mid p_t) && \text{left-corner trans.}\\
&= \mathsf{P}_F(f_t \mid \vec{s}_{t-1}) \cdot && \text{fork (boolean)}\\
&\quad\ \mathsf{P}_J(j_t \mid f_t\, \vec{s}_{t-1}) \cdot && \text{join (boolean)}\\
&\quad\ \mathsf{P}_A(\vec{a}_t \mid f_t\, j_t\, \vec{s}_{t-1}) \cdot && \text{a categories}\\
&\quad\ \mathsf{P}_B(\vec{b}_t \mid f_t\, j_t\, \vec{a}_t\, \vec{s}_{t-1}) \cdot && \text{b categories}\\
&\quad\ \mathsf{P}_P(p_t \mid \vec{b}_t\, \vec{s}_{t-1}) \cdot && \text{part of speech}\\
&\quad\ \mathsf{P}_W(w_t \mid p_t) && \text{word model}
\end{aligned}
$$

But left-corner constraints are more efficient. . .

Left-corner HHMMs are cheaper

In each $\vec{a}_t$, $\vec{b}_t$, only the lowest a_t^d and b_t^d are free (the rest are copied / null):

Left-corner HHMMs are cheaper

In each $\vec{a}_t$, $\vec{b}_t$, only the lowest a_t^d and b_t^d are free (the rest are copied / null):

$$\mathsf{P}_A(\vec{a}_t \mid f_t\, j_t\, \vec{s}_{t-1}) =$$

$$\begin{cases} \prod_{d'=1}^{d-2} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot [\![a_t^{d-1} = a_{t-1}^{d-1}]\!] \cdot \prod_{d'=d}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t = -, j_t = + \\ \prod_{d'=1}^{d-1} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot \mathsf{P}_{-/-}(a_t^d \mid b_{t-1}^{d-1}\, a_{t-1}^d) \cdot \prod_{d'=d+1}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t = -, j_t = - \\ \prod_{d'=1}^{d-1} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot [\![a_t^{d} = a_{t-1}^{d}]\!] \cdot \prod_{d'=d+1}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t = +, j_t = + \\ \prod_{d'=1}^{d} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot \mathsf{P}_{+/-}(a_t^{d+1} \mid b_{t-1}^{d}\, p_{t-1}) \cdot \prod_{d'=d+2}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t = +, j_t = - \end{cases}$$

where $d = \max\{d' | s_{t-1}^{d'} \neq -/-\}$ and $[\![\phi]\!] = 1$ iff ϕ otherwise 0.

Left-corner HHMMs are cheaper

In each $\vec{a}_t$, $\vec{b}_t$, only the lowest a_t^d and b_t^d are free (the rest are copied / null):

$$\mathsf{P}_A(\vec{a}_t \mid f_t\, j_t\, \vec{s}_{t-1}) =$$

$$\begin{cases} \prod_{d'=1}^{d-2} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot [\![a_t^{d-1} = a_{t-1}^{d-1}]\!] \cdot \prod_{d'=d}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t{=}-, j_t{=}+ \\ \prod_{d'=1}^{d-1} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot \mathsf{P}_{-/-}(a_t^d \mid b_{t-1}^{d-1}\, a_{t-1}^d) \cdot \prod_{d'=d+1}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t{=}-, j_t{=}- \\ \prod_{d'=1}^{d-1} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot [\![a_t^{d} = a_{t-1}^{d}]\!] \cdot \prod_{d'=d+1}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t{=}+, j_t{=}+ \\ \prod_{d'=1}^{d} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot \mathsf{P}_{+/-}(a_t^{d+1} \mid b_{t-1}^{d}\, p_{t-1}) \cdot \prod_{d'=d+2}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t{=}+, j_t{=}- \end{cases}$$

$$\mathsf{P}_B(\vec{b}_t \mid f_t\, j_t\, \vec{a}_t\, \vec{s}_{t-1}) =$$

$$\begin{cases} \prod_{d'=1}^{d-2} [\![b_t^{d'} = b_{t-1}^{d'}]\!] \cdot \mathsf{P}_{-/+}(b_t^{d-1} \mid b_{t-1}^{d-1}\, a_{t-1}^d) \cdot \prod_{d'=d}^{D} [\![b_t^{d'} = -]\!], & \text{if } f_t{=}-, j_t{=}+ \\ \prod_{d'=1}^{d-1} [\![b_t^{d'} = b_{t-1}^{d'}]\!] \cdot \mathsf{P}_{-/-}(b_t^d \mid a_t^d\, a_{t-1}^d) \cdot \prod_{d'=d+1}^{D} [\![b_t^{d'} = -]\!], & \text{if } f_t{=}-, j_t{=}- \\ \prod_{d'=1}^{d-1} [\![b_t^{d'} = b_{t-1}^{d'}]\!] \cdot \mathsf{P}_{+/+}(b_t^d \mid b_{t-1}^d\, p_{t-1}) \cdot \prod_{d'=d+1}^{D} [\![b_t^{d'} = -]\!], & \text{if } f_t{=}+, j_t{=}+ \\ \prod_{d'=1}^{d} [\![b_t^{d'} = b_{t-1}^{d'}]\!] \cdot \mathsf{P}_{+/-}(b_t^{d+1} \mid a_t^{d+1}\, p_{t-1}) \cdot \prod_{d'=d+2}^{D} [\![b_t^{d'} = -]\!], & \text{if } f_t{=}+, j_t{=}- \end{cases}$$

where $d = \max\{d' | s_{t-1}^{d'} \neq -/-\}$ and $[\![\phi]\!] = 1$ iff ϕ otherwise 0.

Left-corner HHMMs are cheaper

In each $\vec{a}_t$, $\vec{b}_t$, only the lowest a_t^d and b_t^d are free (the rest are copied / null):

$\mathsf{P}_A(\vec{a}_t \mid f_t\, j_t\, \vec{s}_{t-1}) =$

$$\begin{cases} \prod_{d'=1}^{d-2} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot [\![a_t^{d-1} = a_{t-1}^{d-1}]\!] \cdot \prod_{d'=d}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t{=}-, j_t{=}+ \\ \prod_{d'=1}^{d-1} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot \mathsf{P}_{-/-}(a_t^d \mid b_{t-1}^{d-1}\, a_{t-1}^d) \cdot \prod_{d'=d+1}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t{=}-, j_t{=}- \\ \prod_{d'=1}^{d-1} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot [\![a_t^{d} = a_{t-1}^{d}]\!] \cdot \prod_{d'=d+1}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t{=}+, j_t{=}+ \\ \prod_{d'=1}^{d} [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot \mathsf{P}_{+/-}(a_t^{d+1} \mid b_{t-1}^{d}\, p_{t-1}) \cdot \prod_{d'=d+2}^{D} [\![a_t^{d'} = -]\!], & \text{if } f_t{=}+, j_t{=}- \end{cases}$$

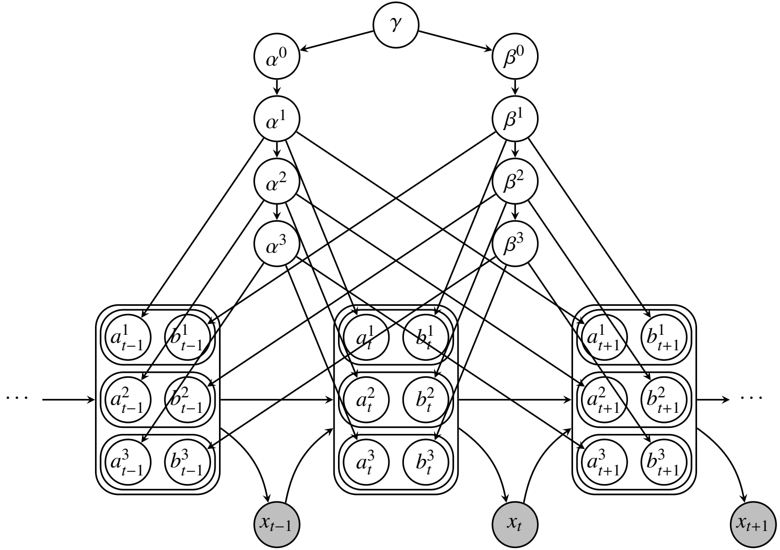
$\mathsf{P}_B(\vec{b}_t \mid f_t\, j_t\, \vec{a}_t\, \vec{s}_{t-1}) =$

$$\begin{cases} \prod_{d'=1}^{d-2} [\![b_t^{d'} = b_{t-1}^{d'}]\!] \cdot \mathsf{P}_{-/+}(b_t^{d-1} \mid b_{t-1}^{d-1}\, a_{t-1}^d) \cdot \prod_{d'=d}^{D} [\![b_t^{d'} = -]\!], & \text{if } f_t{=}-, j_t{=}+ \\ \prod_{d'=1}^{d-1} [\![b_t^{d'} = b_{t-1}^{d'}]\!] \cdot \mathsf{P}_{-/-}(b_t^d \mid a_t^d\, a_{t-1}^d) \cdot \prod_{d'=d+1}^{D} [\![b_t^{d'} = -]\!], & \text{if } f_t{=}-, j_t{=}- \\ \prod_{d'=1}^{d-1} [\![b_t^{d'} = b_{t-1}^{d'}]\!] \cdot \mathsf{P}_{+/+}(b_t^d \mid b_{t-1}^d\, p_{t-1}) \cdot \prod_{d'=d+1}^{D} [\![b_t^{d'} = -]\!], & \text{if } f_t{=}+, j_t{=}+ \\ \prod_{d'=1}^{d} [\![b_t^{d'} = b_{t-1}^{d'}]\!] \cdot \mathsf{P}_{+/-}(b_t^{d+1} \mid a_t^{d+1}\, p_{t-1}) \cdot \prod_{d'=d+2}^{D} [\![b_t^{d'} = -]\!], & \text{if } f_t{=}+, j_t{=}- \end{cases}$$

where $d = \max\{d' | s_{t-1}^{d'} \neq -/-\}$ and $[\![\phi]\!] = 1$ iff ϕ otherwise 0.

So left-corner parser transitions cost only $\mathcal{O}(K^{2\times D} \times 2 \times 2 \times K^2)$.

Labels can be unified using priors across depth levels



Bibliography I

Abney, S. P., & Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *J. Psycholinguistic Research*, *20*(3), 233–250.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Unpublished doctoral dissertation, Carnegie Mellon.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA, USA: Harvard University Press.

Karlsson, F. (2007). Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, *43*, 365–392.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Miller, G. A., & Isard, S. (1964). Free recall of self-embedded english sentences. *Information and Control*, *7*, 292–303.

Bibliography II

Resnik, P. (1992). Left-corner parsing and psychological plausibility. In *Proceedings of COLING* (pp. 191–197). Nantes, France.

Schuler, W., AbdelRahman, S., Miller, T., & Schwartz, L. (2010). Broad-coverage incremental parsing using human-like memory constraints. *Computational Linguistics*, *36*(1), 1–30.

Stabler, E. (1994). The finite connectivity of linguistic structure. In *Perspectives on sentence processing* (pp. 303–336). Lawrence Erlbaum.

van Schijndel, M., Murphy, B., & Schuler, W. (2015). Evidence of syntactic working memory usage in MEG data. In *Proceedings of CMCL 2015.* Association for Computational Linguistics.

van Schijndel, M., Nguyen, L., & Schuler, W. (2013). An analysis of memory-based processing costs using incremental deep syntactic dependency parsing. In *Proc. of CMCL 2013.* Association for Computational Linguistics.

Bibliography III

Weiss, S., Mueller, H. M., Schack, B., King, J. W., Kutas, M., & Rappelsberger, P. (2005). Increased neuronal communication accompanying sentence comprehension. *International Journal of Psychophysiology*, *57*, 129–141.