An Analysis of Frequency- and Memory-Based Processing Costs

Marten van Schijndel William Schuler
Department of Linguistics
The Ohio State University

June 10, 2012

Motivation

Observation isn't explanation

Many current metrics predict complexity with no cognitive explanation.

- Surprisal and entropy reduction reflect corpus statistics.

Motivation

Observation isn't explanation

Many current metrics predict complexity with no cognitive explanation.

- Surprisal and entropy reduction reflect corpus statistics.

Goal: An Explanation

- How do current theories of working memory fit with current theories of language processing?
- Do memory effects predict difficulty over frequency effects?
- Provide a rationale for *why* humans have certain difficulties

Overview

Hypothesis

Memory effects cause processing difficulty beyond frequency effects

Overview

Hypothesis

Memory effects cause processing difficulty beyond frequency effects

1. Working memory primer
2. Memory and language processing theories
3. Introduce connected component parser
4. Eye-tracking evaluation
5. Results

Working Memory

Temporal and Sequential Cueing

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

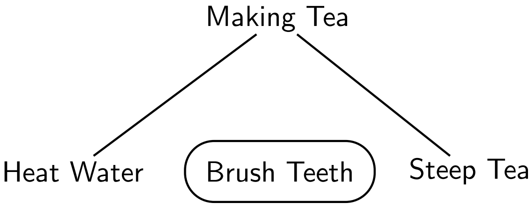
- Learned *sequential* associations
- Contextual *temporal* associations

Working Memory

Temporal and Sequential Cueing

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations



Temporal Cueing in the Morning

Working Memory

Temporal and Sequential Cueing

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations

Focus

Attended vs Passive States [McElree, 2006]

Working Memory

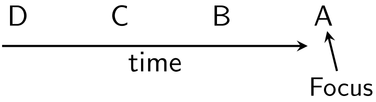
Temporal and Sequential Cueing

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations

Focus

Attended vs Passive States [McElree, 2006]



Working Memory

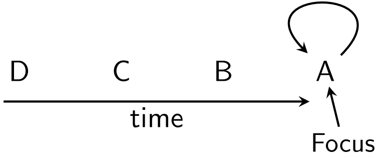
Temporal and Sequential Cueing

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations

Focus

Attended vs Passive States [McElree, 2006]



Working Memory

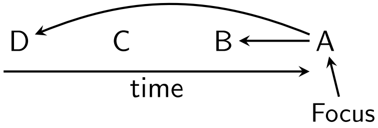
Temporal and Sequential Cueing

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations

Focus

Attended vs Passive States [McElree, 2006]



Working Memory

Temporal and Sequential Cueing

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations

Focus

Attended vs Passive States [McElree, 2006]

$$\textbf{Difficulty with}\begin{cases}\text{Temporal cueing}\\ \text{(Accessing non-focused information)}\end{cases}$$

$$\textbf{Temporal cueing}\begin{cases}\text{Resolving embedded dependencies}\end{cases}$$

Key: Inhibition Facilitation

Language Processing

Dependency Locality Theory [Gibson, 2000]

$$\textbf{Difficulty with} \begin{cases} \text{Unresolved dependencies} \end{cases}$$

$$\textcolor{red}{\textbf{Storage cost}} \begin{cases} \text{Beginning dependencies} \\ \text{Maintaining dependencies} \end{cases}$$

$$\textcolor{red}{\textbf{Integration cost}} \begin{cases} \text{Resolving dependencies} \end{cases}$$

Language Processing

ACT-R [Lewis et al., 2006]

$$\textbf{Difficulty with} \begin{cases} \text{Activation decay} \\ \text{Similarity interference} \end{cases}$$

$$\textbf{Encoding cost} \begin{cases} \text{Beginning a new dependency} \end{cases}$$

$$\textbf{Retrieval cost} \begin{cases} \text{Resolving a dependency} \end{cases}$$

Retrieval can be *facilitated* by re-activations.

Language Processing

Dynamic Recruitment [Just and Varma, 2007]
Difficult constructions → extra processing resources

Difficulty with { Center embeddings

Recruitment { Beginning embeddings

Release { Completing embeddings

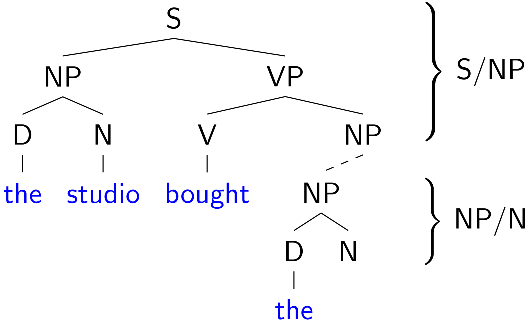
Language Processing

Embedding Difference [Wu et al., 2010]

Increased embedding depth $\Big\{$ Beginning embeddings

Reduced embedding depth $\Big\{$ Completing embeddings

Connected Components



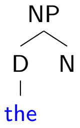
'S/NP' and 'NP/N' represent unresolved dependencies

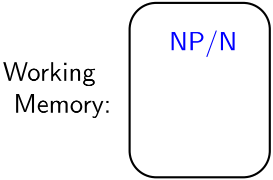
Predictions

Theory	Encoding	Integration
Hier. Sequential Prediction		**positive**
Dependency Locality Theory	**positive**	**positive**
ACT-R	**positive**	**positive**
Dynamic Recruitment	**positive**	**negative**
Embedding Difference	**positive**	**negative**

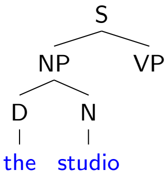
Predicted correlation of parse operations to reading times under each theory

Connected Component Parsing

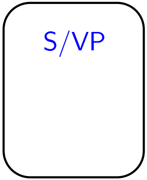




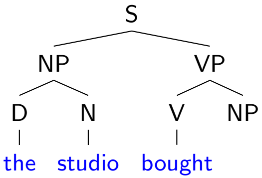
Connected Component Parsing



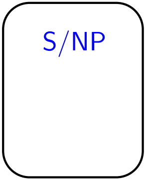
Working Memory:



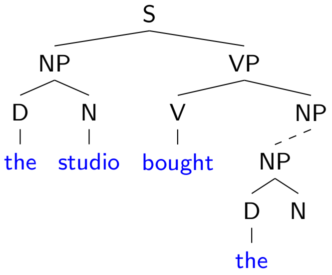
Connected Component Parsing



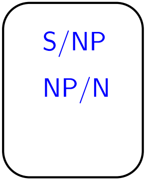
Working Memory:



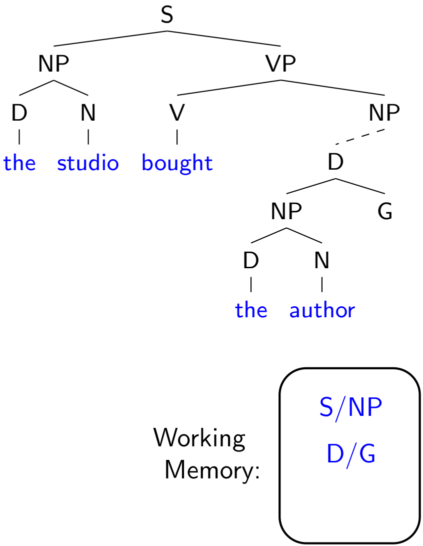
Connected Component Parsing



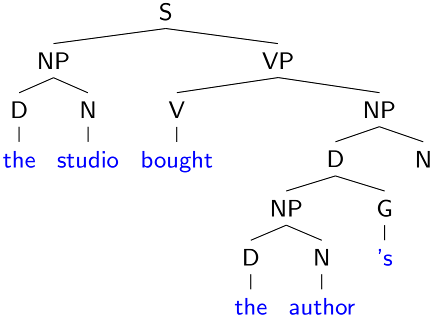
Working Memory:



Connected Component Parsing



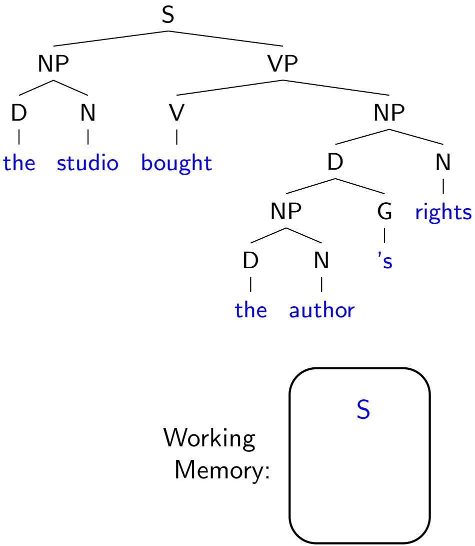
Connected Component Parsing



Working Memory:

S/N

Connected Component Parsing



Parser Operations

F and L binary decisions (+,–) made at each timestep

- **F(irst)**: Current word is the **first** element of a new embedding
- **L(ast)**: Current word is the **last** element of an embedding

Only one F, only one L [van Schijndel et al, 2013]

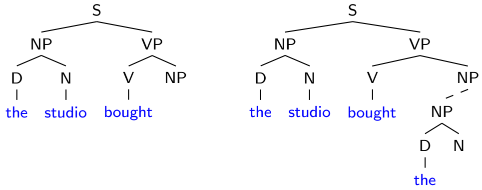
Parser Operations

F and L binary decisions (+,–) made at each timestep

- **F(irst)**: Current word is the **first** element of a new embedding
- **L(ast)**: Current word is the **last** element of an embedding

Only one F, only one L [van Schijndel et al, 2013]

- F+L– (Encode): Create a new connected component



Encode

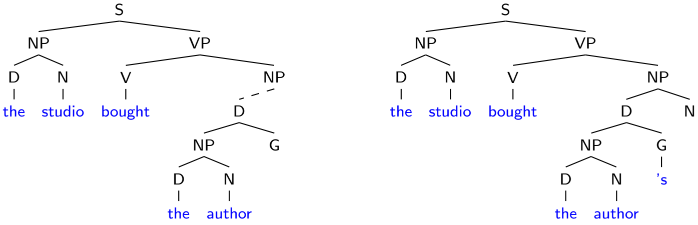
Parser Operations

F and L binary decisions (+,–) made at each timestep

- **F(irst)**: Current word is the **first** element of a new embedding
- **L(ast)**: Current word is the **last** element of an embedding

Only one F, only one L [van Schijndel et al, 2013]

- F+L– (Encode): Create a new connected component
- F–L+ (Integrate): Combine two connected components



Integrate

Eye Tracking

- Assumption: Slower reading = difficulty
- How much can be processed up to a given point?
- Many different metrics (fixation duration, regression, etc)

Eye Tracking

- Assumption: Slower reading = difficulty
- How much can be processed up to a given point?
- Many different metrics (fixation duration, regression, etc)

Measure of choice: Go-Past Duration [Clifton et al., 2007]

Eye Tracking

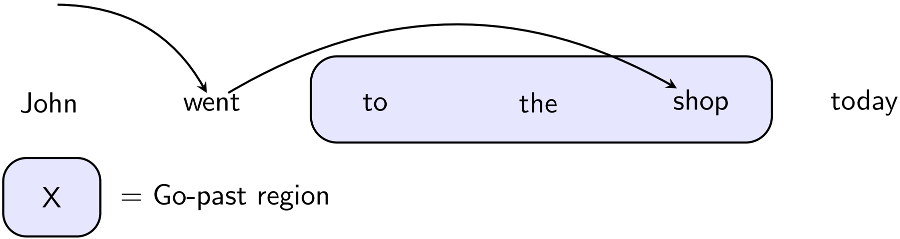
Go-past durations:

John went to the shop today

Cumulative factors are summed over the go-past region
Non-cumulative factors are based on the initial word in a region (shop)

Eye Tracking

Go-past durations:



Cumulative factors are summed over the go-past region

Non-cumulative factors are based on the initial word in a region (shop)

Training

Parser accuracy is comparable to Berkeley [van Schijndel et al., 2012]

- Parser and Lexicon: WSJ02-21 [Marcus et al., 1993]
  - 39,832 sentences
  - 950,028 words
- Ngrams: Brown [Francis and Kucera, 1979], WSJ02-21, BNC, Dundee[Kennedy et al., 2003]
  - 5,052,904 sentences
  - 87,302,312 words

Ngrams calculated using SRILM [Stolcke, 2002] with modified Kneser-Ney smoothing [Chen and Goodman, 1998]

Evaluation

- Dundee corpus [Kennedy et al., 2003]
  - 10 subjects
  - 2,388 sentences
  - 58,439 words
  - 260,124 go-past durations
- Filtered Dundee corpus
  - 154,168 go-past durations

Exclusions: UNK-threshold 5, first and last of a line, fixations skipping more than 4 words (track/attention loss)

Metric Calculations: Probability-weighted, parallel model

Baseline Metrics

Fitting a linear mixed effects model (*lmer* in R)

Fixed Effects

- Word length
- Sentence position
- Prev, Next word fixated?
- Unigram and bigram probs
- Surprisal
- Region length
- Cumulative surprisal
- Cumulative entropy reduction
- Joint interactions
- Spillover predictors

By-subject random slopes (Note: Not in paper)

- Effect of interest (e.g. Encode)
- Prev word fixated?
- Cumulative surprisal
- Region length

With Subject and Item random intercepts
Fit to log-transformed durations

Predictions - Revisited

Theory	Encoding	Integration
Hier. Sequential Prediction		**positive**
Dependency Locality Theory	**positive**	**positive**
ACT-R	**positive**	**positive**
Dynamic Recruitment	**positive**	**negative**
Embedding Difference	**positive**	**negative**

Predicted correlation of parse operations to reading times under each theory

Results

Operation	Factor	Coeff	Std. Error	t-score	p-value
Encoding	F+L–	0.023	0.005	4.238	0.001
Integration	F–L+	-0.015	0.005	-3.215	0.007
Cue Active	F–L–	0.002	0.003	0.800	0.437
Cue Awaited	F+L+	-0.004	0.003	-1.298	0.22

Significance of Improvement over Baseline

Each FL factor is cumulative

Conclusion

- No positive integration cost with frequency

Conclusion

- No positive integration cost with frequency
- Significant negative integration cost

Conclusion

- No positive integration cost with frequency
- Significant negative integration cost
- Supports: Dynamic Recruitment, Embedding Difference

Conclusion

- No positive integration cost with frequency
- Significant negative integration cost
- Supports: Dynamic Recruitment, Embedding Difference

- No evidence of DLT's maintenance cost

Conclusion

- No positive integration cost with frequency
- Significant negative integration cost
- Supports: Dynamic Recruitment, Embedding Difference

- No evidence of DLT's maintenance cost
- Confounds assumption of Slow = Difficult

Conclusion

- No positive integration cost with frequency
- Significant negative integration cost
- Supports: Dynamic Recruitment, Embedding Difference

- No evidence of DLT's maintenance cost
- Confounds assumption of Slow = Difficult
- Remaining inhibition suggests difficulty beyond frequency effects (perhaps a cause of frequency effects)

Fin

Thanks!

Thanks to Kodi Weatherholtz and Rory Turnbull for their assistance with R-wrangling and working with linear mixed effect models!

Thanks to Peter Culicover, Micha Elsner, and the OSU CompLing group for feedback on the project.

Questions?

Frequency Effects

Surprisal [Hale, 2001]

Predictability of a word given the context:

$$surprisal(x_t) = -\log_2 \left(\frac{\sum_{s \in S(x_1 \ldots x_t)} P(s)}{\sum_{s \in S(x_1 \ldots x_{t-1})} P(s)} \right) \tag{1}$$

Entropy Reduction [Hale, 2003]

Entropy is a measure of uncertainty:

$$H(x_{1 \ldots t}) = \sum_{s \in S(x_1 \ldots x_t)} -P(s) \cdot \log_2 P(s) \tag{2}$$

The reduction in uncertainty caused by observing x_t:

$$\Delta H(x_{1 \ldots t}) = \max(0, H(x_{1 \ldots t-1}) - H(x_{1 \ldots t})) \tag{3}$$

$S(x_1 \ldots x_t)$ = trees whose leaves have $x_1 \ldots x_t$ as a prefix

Eye Tracking

Go-past durations:

John went to the shop today

Cumulative factors are summed over the go-past region
Non-cumulative factors are based on the initial word in a region (shop)

Eye Tracking

Go-past durations:

John went to the shop today

X = Go-past region

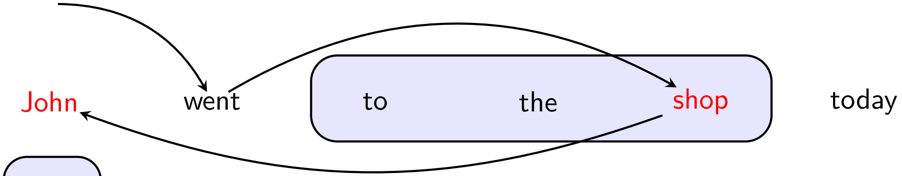
Red = Fixation in go-past duration

Cumulative factors are summed over the go-past region

Non-cumulative factors are based on the initial word in a region (shop)

Eye Tracking

Go-past durations:



X = Go-past region

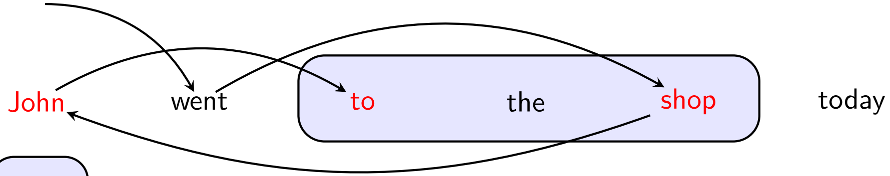
Red = Fixation in go-past duration

Cumulative factors are summed over the go-past region

Non-cumulative factors are based on the initial word in a region (shop)

Eye Tracking

Go-past durations:



X = Go-past region

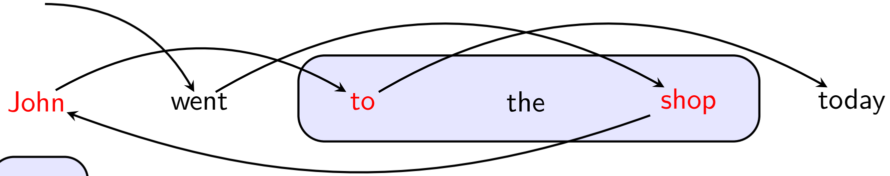
Red = Fixation in go-past duration

Cumulative factors are summed over the go-past region

Non-cumulative factors are based on the initial word in a region (shop)

Eye Tracking

Go-past durations:



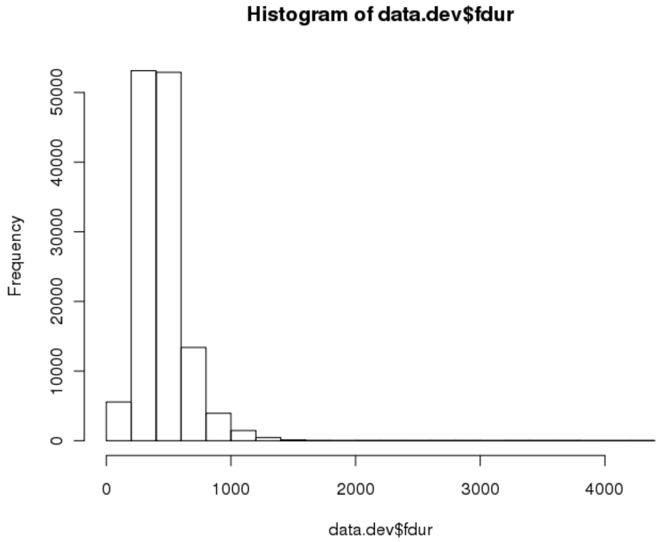
X = Go-past region

Red = Fixation in go-past duration

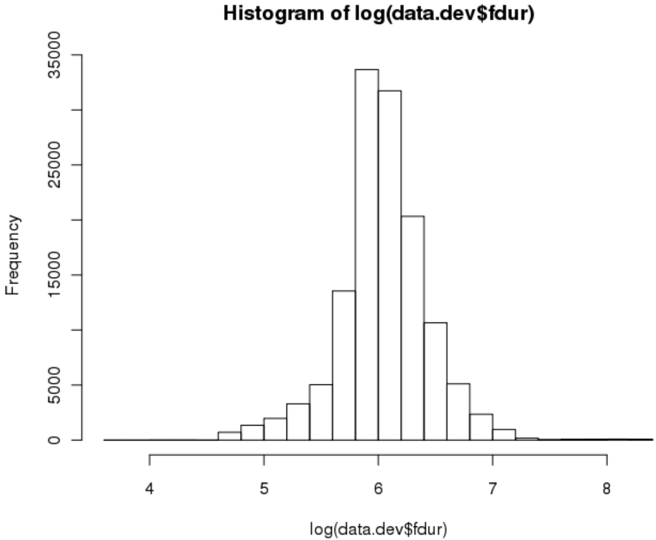
Cumulative factors are summed over the go-past region

Non-cumulative factors are based on the initial word in a region (shop)

Transforming the response variable



Transforming the response variable



Bibliography I

Botvinick, M. (2007).
Multilevel structure in behavior and in the brain: a computational model of Fuster's hierarchy.
Philosophical Transactions of the Royal Society, Series B: Biological Sciences, 362:1615–1626.

Chen, S. F. and Goodman, J. (1998).
An empirical study of smoothing techniques for language modeling.
Technical report, Harvard University.

Clifton, C., Staub, A., and Rayner, K. (2007).
Eye movements in reading words and sentences.
In *Eye movements: A window on mind and brain*, pages 341–372. Elsevier.

Francis, W. N. and Kucera, H. (1979).
The brown corpus: A standard corpus of present-day edited american english.

Bibliography II

Gibson, E. (2000).
The dependency locality theory: A distance-based theory of linguistic complexity.
In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126, Cambridge, MA. MIT Press.

Hale, J. (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*, pages 159–166, Pittsburgh, PA.

Hale, J. (2003).
Grammar, Uncertainty and Sentence Processing.
PhD thesis, Cognitive Science, The Johns Hopkins University.

Bibliography III

Howard, M. W. and Kahana, M. J. (2002).
A distributed representation of temporal context.
Journal of Mathematical Psychology, 45:269–299.

Just, M. A. and Varma, S. (2007).
The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition.
Cognitive, Affective, & Behavioral Neuroscience, 7:153–191.

Kennedy, A., Pynte, J., and Hill, R. (2003).
The Dundee corpus.
In *Proceedings of the 12th European conference on eye movement*.

Lewis, R. L., Vasishth, S., and Dyke, J. A. V. (2006).
Computational principles of working memory in sentence comprehension.
Trends in Cognitive Science, 10(10):447–454.

Bibliography IV

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993).
Building a large annotated corpus of English: the Penn Treebank.
Computational Linguistics, 19(2):313–330.

McElree, B. (2006).
Accessing recent events.
The Psychology of Learning and Motivation, 46:155–200.

Roark, B. (2001).
Probabilistic top-down parsing and language modeling.
Computational Linguistics, 27(2):249–276.

Schuler, W. (2009).
Parsing with a bounded stack using a model-based right-corner transform.
In *Proceedings of NAACL/HLT 2009*, NAACL '09, pages 344–352, Boulder, Colorado. Association for Computational Linguistics.

Bibliography V

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

Stolcke, A. (2002).
Srilm – an extensible language modeling toolkit.
In *Seventh International Conference on Spoken Language Processing*.

van Schijndel, M., Exley, A., and Schuler, W. (2012).
Connectionist-inspired incremental PCFG parsing.
In *Proceedings of CMCL 2012*. Association for Computational Linguistics.

van Schijndel, M., Exley, A., and Schuler, W. (in press).
A model of language processing as hierarchic sequential prediction.
Topics in Cognitive Science.

Bibliography VI

Wu, S., Bachrach, A., Cardenas, C., and Schuler, W. (2010).
Complexity metrics in an incremental right-corner parser.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 1189–1198.