Integrating Surprisal and Dependency Locality Theory: A Broad Coverage Model

Vera Demberg and Frank Keller
University of Edinburgh
v.demberg@ed.ac.uk, keller@inf.ed.ac.uk

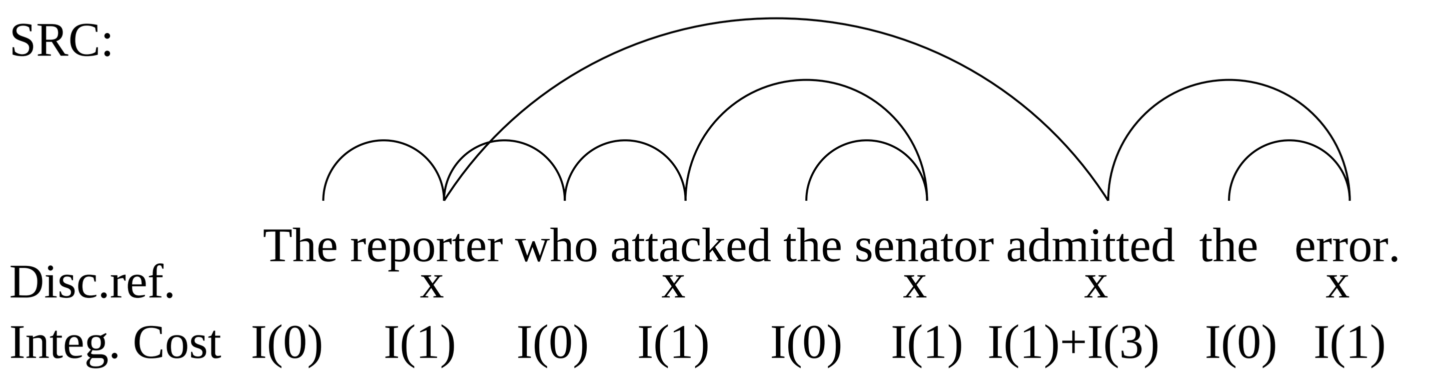
Introduction

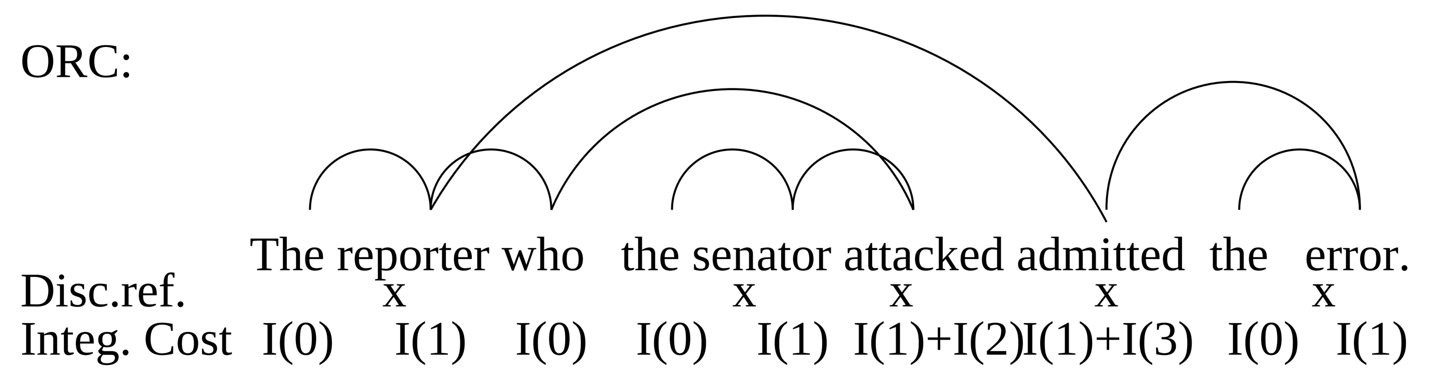
- **Dependency Locality Theory** (DLT; Gibson, 1998) predicts processing difficulty based on integration cost (IC), the effort required when integrating a head with its dependents.
- Alternative: **Surprisal** (Hale, 2001) relates processing difficulty to the amount of information processed at each word in the sentence.
- Demberg and Keller (2007) tested both IC and Surprisal on an eyetracking corpus, and found that Surprisal, but not IC, predicts reading times in a broad-coverage setting.
- Here we propose a variant of DLT capable of making **broad coverage predictions** by redefining integration cost in terms of surprisal.

Integration Cost

- Integration cost occurs at heads of phrases when dependents are integrated; the number of discourse referents crossed determines cost.
- We used Minipar to compute dependency structures, determine discourse referents, and compute the integration cost for every word.

Example:





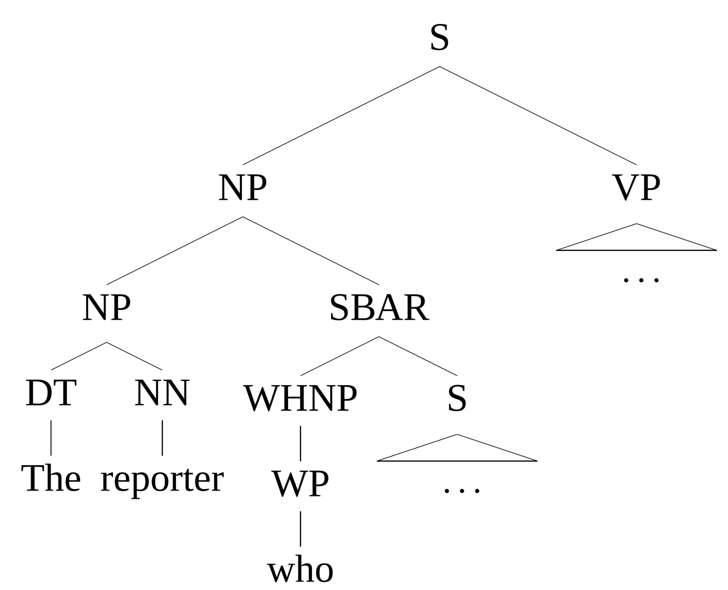
Surprisal

- Processing difficulty is triggered when surprising linguistic material is processed (structures that are unlikely given the current input).
- The Surprisal S_{k+1} at word w_{k+1} is defined as the conditional probability of w_{k+1} given the left context (words w_1 to w_k):

$$\begin{aligned} S_{k+1} &= -\log P(w_{k+1}|w_1 \cdots w_k) \\ &= \log P(w_1 \cdots w_k) - \log P(w_1 \cdots w_{k+1}) \end{aligned} \quad (1)$$

  where $P(w_1 \cdots w_k)$ is the prefix probability of the string $w_1 \cdots w_k$.
- We used Roark's (2001) incremental lexicalized parser to compute prefix probabilities.

Example:



Example	Rule	Rule prob.
The reporter who ...	S → VP NP	$p = 0.6$
The reporter who ...	NP → NP SBAR	$p = 0.004$
The reporter	NP → DT NN	$p = 0.5$
reporter	NN → reporter	$p = 0.0002$
who	WHNP → WP	$p = 0.2$
who	WP → who	$p = 0.8$

Surprisal-based Integration Cost

- Surprisal-based Integration Cost (SIC) combines DLT and Surprisal.
- At a head w_h, the SIC of a syntactic dependent w_d is defined as $S_{d...h}$, the cumulative Surprisal of the words intervening between w_h and w_d.
- The Surprisal of a region of words can be defined as:

$$S_{d...h} = \log P(w_1 \cdots w_h) - \log P(w_1 \cdots w_d) \quad (2)$$

- If d directly follows h, then SIC reduces to standard Surprisal.
- The total SIC at h is defined as the sum of the Surprisal of all its dependents:

$$SIC_h = \sum_d S_{d...h} \quad (3)$$

- SIC is distance-based like IC, but distance is measured as Surprisal of the intervening words, not as number of intervening discourse referents.
- **Hypothesis:** intervening high-Surprisal material makes it more difficult to keep a dependent in memory.

Experiments

Method

- Data: Dundee Corpus (Kennedy & Pynte, 2005); contains eyetracking record of 10 subjects reading 51,000 words of newspaper text.
- Method: hierarchical mixed effects model with reading time as the dependent variable and SIC, standard IC, or Surprisal as target variable.
- The models included nine control variables that are known to influence reading time (e.g., frequency, launch distance).

Results and Discussion

- The results we report are for the **verbs and nouns** in the corpus (the other words don't have any dependents and therefore zero SIC); model $R^2 = 0.166$.
- **Surprisal** not a significant predictor of reading times (unlike on the full Dundee corpus).
- **Standard Integration Cost** significant negative predictor, contrary to theoretical expectation (but consistent with Demberg & Keller, 2007).
- **Surprisal-based Integration Cost** significant positive predictor.
- Correlations: IC and SIC uncorrelated with Surprisal ($r = 0.08$), but IC and SIC significantly correlated ($r = 0.69$). But no colinearity (tested using scaling, VIF): IC and SIC explain separate portions of the variance.

Predictor	(interc)	word length	word freq.	landing pos.	prev. word fixated	bigram prob.	SIC	length : freq.	length : wlpos
Coefficient	126.36	29.71	15.49	-1.48	-18.61	-2.11	0.04	-4.71	-1.20
p-value	***	***	***	**	***	***	**	***	***

Conclusions

- Surprisal-based Integration Cost predicts reading times in natural, contextualized text.
- A realistic model of processing difficulty needs to integrate both locality-based factors (integration cost) and probabilistic measures (Surprisal).

References

Demberg, V., & Keller, F. 2007. Do Theories of Syntactic Processing Difficulty Scale up to Naturally Occuring Text?Do theories of syntactic processing difficulty scale up to naturally occuring text? Talk at the 13th Conference on Architectures and Mechanisms for Language Processing, Turku.

Gibson, E. 1998. Linguistic complexity: locality of syntactic dependenciesLinguistic complexity: locality of syntactic dependencies. Cognition681-76.

Hale, J. 2001. A Probabilistic Earley Parser as a Psycholinguistic ModelA probabilistic earley parser as a psycholinguistic model. In Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational LinguisticsProceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics (Vol. 2, pp. 159–166). Pittsburgh, PA.

Kennedy, A., & Pynte, J. 2005. Parafoveal-on-foveal Effects in Normal ReadingParafoveal-on-foveal effects in normal reading. Vision Research45153–168.

Roark, B. 2001. Probabilistic Top-Down Parsing and Language ModelingProbabilistic top-down parsing and language modeling. Computational Linguistics272249–276.

School of Informatics