A Probabilistic Corpus-Based Model of Syntactic Parallelism

Amit Dubey
(Joint work with Frank Keller and Patrick Sturt)

Human Communications Research Centre
University of Edinburgh

Eurocogsci 2007
Probabilistic Approach to Computational Psycholinguistics

Long-term goals of computational psycholinguistics:

- Build cognitively plausible computational models of human language processing
- Models should mimic the accuracy, coverage and efficiency of the human language processor

Agenda behind this research: achieve this using probabilistic models from computational linguistics
Among the benefits of this approach is principled way to deal with ambiguity

However, this talk is not about ambiguity…

…but rather about a process which makes unambiguous structures easier to process
Syntactic Parallelism in Coordination

```
S
  NP  VP
    NP  NP
      V  Det  Adj  N
      wrote  a long novel  and
      CC  N
      a short poem
```
Introduction
Corpus Studies
Model Based on Corpus Studies
Model Based on ACT-R
Experimental Setting
Conclusions

Syntactic Parallelism in Coordination

Terry wrote a long novel and a short poem

Modelling Syntactic Parallelism
Syntactic Parallelism in Coordination

Terry wrote a long novel and a short poem.
Introduction
Corpus Studies
Model Based on Corpus Studies
Model Based on ACT-R
Experimental Setting
Conclusions

Syntactic Parallelism in Coordination

S

NP

VP

N

V

long and

Terry wrote a novel

CC

Parallel Structure

Det Adj N

NP

Det Adj N

NP

a short poem

Dubey
Modelling Syntactic Parallelism 4
Syntactic Parallelism in Coordination

Introduction
Corpus Studies
Model Based on Corpus Studies
Model Based on ACT-R
Experimental Setting
Conclusions

Dubey
Modelling Syntactic Parallelism

Terry wrote a long novel and a short poem.
Syntactic Parallelism in Coordination

Terry wrote a long novel and a short poem.

Parallel Structure

Speed-up
Syntactic Parallelism in Coordination

S

NP

NP

NP

NP

N               V

Det               a

N               novel

CC        and

Det               Adj

N               poem

Terry wrote a novel and Terry wrote a short poem.
Syntactic Parallelism in Coordination

Dubey Modelling Syntactic Parallelism
Syntactic Parallelism in Coordination

S

NP
Terry wrote

VP

NP
Det
N
a novel

NP
CC
and

NP
Det
Adj
a short poem
Syntactic Parallelism in Coordination

Dubey
Modelling Syntactic Parallelism
Syntactic Parallelism in Coordination

S
  NP
  V

Terry wrote

NP
  Det
  a

novel

CC

and

NP
  Det
  a

short

Adj

poem

Not Parallel
Syntactic Parallelism in Coordination

Introduction
Corpus Studies
Model Based on Corpus Studies
Model Based on ACT-R
Experimental Setting
Conclusions

Terry wrote a novel and Terry wrote a short poem.

Not Parallel
No Speed-up

S
VPNP
N V
NP
Det N
NP NP
Det Adj N

Dubey Modelling Syntactic Parallelism 4
Research Question

- This effect is well-established (cf. Frazier et al., 2000; Frazier et al., 1984; Carlson, 2002; Mauner et al., 1995)
- Frazier et al. (2000) and Apel et al. (2007) argue the parallelism effect is due to a specialised mechanism
- Can the parallelism effect be explained using a general mechanism: syntactic priming?
  - Begin with a series of corpus studies
  - Move on to several modeling studies
  - How the models relate to new experimental studies
What to Count?

- Pursue a different approach than previous corpus studies of priming (e.g. Gries, 2005; Szmrecsanyi 2005)
- More amenable to modeling studies: *probabilistic approach*
- Pick a particular rule
- Pick a *priming* region where the rule is (potentially) used a first time
- Pick a *target* region where the rule is (potentially) reused
- We may vary the grain size (i.e. the size of the prime and target region)
In the case of coordination, the first conjunct is the prime, and the second the target.

In the case of coordination, the first conjunct is the prime, and the second the target.

Terry wrote a long novel and Terry wrote a short poem.
Level of Granularity

Coordination

Terry wrote a short poem and a long novel.

Within sentences

Terry wrote a short poem for a long novel.

Between sentences

Terry wrote a short poem. His publisher wanted a long novel.
Prior and Adaptation

- Pick one rule
- Prior probability
  - Relative frequency of this rule occurring in a corpus (whether or not it was reused)
  - \[ P(\text{prior}) = \frac{\text{count}(\text{structure occurs in target region})}{N} \]
- Adaptation probability
  - Relative frequency of parallel structures: i.e. relative frequency of this rule occurring \textbf{given} that it already has occurred in the prime
  - \[ P(\text{adapt}) = \frac{\text{count}(\text{parallel})}{\text{count}(\text{structure occurs in prime region})} \]
- Evidence for priming (tendency to reuse structures) if adaptation > prior
How to Count?

- One corpus study for each level of granularity
- Focus on several kinds of $NP$
  - $NP \rightarrow NP\ SBAR$, $NP \rightarrow NP\ PP$, $NP \rightarrow N$, etc.
- Found statistically significant effects in (almost) all cases, in several different parsed corpora
Brown Corpus: Coordination

Adaptation within coordinate structures in the Brown corpus
Brown Corpus: Within sentences

Adaptation within sentences in the Brown corpus
Introduction

Corpus Studies
Model Based on Corpus Studies
Model Based on ACT-R
Experimental Setting
Conclusions

Brown Corpus: Between sentences

Adaptation between sentences in the Brown corpus
Adaptation within documents in the Brown corpus (all items exhibit weak yet statistically significant positive adaptation)
Results

- Results of corpus study
  - Strong effect in coordination
  - But weaker effect is prevalent elsewhere (experimental work suggests this shouldn’t happen)
- For further details: see Dubey et al. (2005) at HLT/EMNLP 2005
  - Includes results for other corpora, including spontaneous speech
Introducing the Model

- General framework: PCFG-based parser (e.g. Jurafsky, 1996; Crocker and Brants, 2000)
- Broad-coverage, incremental
- PCFG: Context-free rules + probabilities
Return to question: is there a specialised mechanism for coordination?

Test a specialised model vs. a general model

**General model**  Within-sentence Adaptation, similar to corpus study

**Specific model**  Copying: Copy-\( \alpha \) of Frazier and Clifton (2001); C-Module of Apel et al. (2007)
Copy Model

Terry wrote a novel and a book

NP

NP
Copy Model

Terry wrote a novel and a book
Copy Model

Terry wrote a novel and a book

Terry wrote a novel and a book
How does the computational model relate to reading-time predictions?

- Use sentence probabilities
  - The higher the probability, the easier to read

- There are other proposals (which are likely highly correlated with whole-sentence probabilities) which can’t be used
  - The method of Jurafsky (1996) requires ambiguity
  - The information-theoretic methods of Hale (2003) and Hale (2001) make word-by-word predictions, but we are interested in regions
Experimental Items

- Specialised mechanism is argued in Frazier et al. (2000)
- Copy and Adaptation parser model the specialised and general hypotheses, respectively
- Run parsers on experimental items from Frazier et al. (2000), Experiment 3
- Terry wrote \{ a long novel, a novel \} and a short poem
The Copy Model successfully models experimental data of Frazier et al. (2000) (Wilcoxon signed rank test: $N = 24$; $Z = 4.27; p < .001$)

However, the more general Adaptation Model also models the results ($Z = 1.67; p < .05$)

From this, may we conclude a specialised mechanism isn’t necessary to model the experimental results?

Modeling studies previously appeared in Dubey et al. (2006) at COLING/ACL
- Adaptation model is somewhat ad-hoc
- We must pick grain size, and ‘prime’ is binary
- Would expect time-course effects, e.g. decay
- It is possible to model such a decay
- Base on practise effects of ACT-R models (Anderson et al., 2004; Lovett, 1998)
A Model Based on Decay

- Adapt some aspects of ACT-R into a probabilistic system
- Key idea: define PCFG system in terms of ACT-R concepts
- Abstract away from low-level details of ACT-R
- When rule is encountered at time $t$
  - Add $t^{-d}$ to numerator/denominator
- Set $d = 0.5$, following Lewis and Vasishth (2005)
Recall the definition of PCFG probability:

$$P(\beta|A) = \frac{f(A \rightarrow \beta)}{f(A)}$$

The definition of production rule probability with decay:

$$P = \frac{\sum_{i \in \text{Successes}} t_i^d}{\sum_{i \in \text{Successes}} t_i^d + \sum_{j \in \text{Failures}} t_j^d}$$

These can be combined to:

$$P(\beta|A) = \frac{\sum_{i \in A \rightarrow \beta} t_i^d}{\sum_{i \in A \rightarrow \beta} t_i^d + \sum_{j \in A \rightarrow \gamma, \gamma \neq \beta} t_j^d}$$
Counting probabilities: PCFG

probability of \( NP \rightarrow \text{Det Adj N} \)

\[
\frac{1}{1} = 100\%
\]

The new car is red.

Go to school!

The nice man seated at the table likes seafood.

Terry wrote a long novel and a short poem

Dubey Modelling Syntactic Parallelism 25
Counting probabilities: PCFG

probability of $NP \rightarrow \text{Det Adj N}$

$\frac{1}{1} = 100\%$

$\frac{1}{1+1} = 50\%$
Counting probabilities: PCFG

probability of $NP \rightarrow Det \ Adj \ N$

\[
\frac{1}{1} = 100\%
\]

\[
\frac{1}{1+1} = 50\%
\]

\[
\frac{1+1}{1+1+1+1+1} = 33\%
\]
Counting probabilities: PCFG

probability of
\[ NP \rightarrow \text{Det Adj N} \]

\[ \frac{1}{1} = 100\% \]
\[ \frac{1}{1+1} = 50\% \]
\[ \frac{1+1}{1+1+1+1+1+1+1+1} = 33\% \]
\[ \frac{1+1+1+1}{1+1+1+1+1+1+1+1+1+1+1+1} = 44\% \]
Counting probabilities: Decay Model

probability of \( NP \rightarrow Det \ Adj \ N \)

\[
\frac{500^{-0.5}}{500^{-0.5}} = 100\%
\]

The new car is red.

Go to school!

The nice man seated at the table likes seafood.

Terry wrote a long novel and a short poem.
Counting probabilities: Decay Model

probability of $NP \rightarrow Det\ Adj\ N$

$$\frac{500^{-0.5}}{500^{-0.5}} = 100\%$$

$$\frac{2000^{-0.5} + 100^{-0.5}}{2000^{-0.5} + 100^{-0.5}} = 18\%$$
The Decay Model also predicts a statistically significant speedup in the 2nd conjunct (but effect is smaller than the Adapt or Copy models)

However, has several advantages:
- More plausible (temporal decay)
- No need to set grain size
- ‘Closer’ to prior PCFG distribution than Adapt model

From this point on, I will only discuss the Decay Model
How General is the Model?

- Knöferle and Crocker (2006): parallelism in German word order
  - Different construction (NP and NP vs. S and S)
  - Different manipulation (sentential word order)
  - Different language

- Is it possible to model this very different data set?
  - Yes, all that is necessary is to train on a German corpus
  - Notably, nothing else changed: the model was *exactly* the same one used in the previous experiment
Several minutes ago, the violinist praised the singer.
Several minutes ago, the violinist praised the singer
Several minutes ago, the violinist praised the singer.

Knöferle & Crocker (2006)
Vor ein paar Minuten lobte der Geiger den Sänger und in diesem Augenblick preist der Trommler den Dichter.

Several minutes ago, the violinist complimented the singer and at this moment the drummer is praising the poet.

Vor ein paar Minuten lobte den Sänger der Geiger und in diesem Augenblick preist der Trommler den Dichter.

Several minutes ago, the violinist complimented the singer and at this moment the drummer is praising the poet.
Frazier et al. (2000) found no parallelism effect outside coordinated context:

(1) a. The nurse and the shy patient...
   b. The nurse saw the shy patient...

- Apel et al. (2007) replicated this in German (fixing some methodological problems)
- But our model predicts facilitation... or does it? Depends upon the grammar
Frazier et al. (2000) found no parallelism effect outside coordinated context:

(1)  

a. The young nurse and the shy patient. . .
b. The young nurse saw the shy patient. . .

- Apel et al. (2007) replicated this in German (fixing some methodological problems)
- But our model predicts facilitation. . . or does it? Depends upon the grammar
With a grammar which includes grammatical relations, the model makes the correct prediction.

\[
\text{NP} \left[ \text{GF SUBJ} \right] \rightarrow \text{Det Adj N}
\]

\[
\text{NP} \left[ \text{GF OBJ} \right] \rightarrow \text{Det Adj N}
\]

Which features should be included?
No coordination, same grammatical function:

(2)  
  a. The nurse and the bitter pill killed Sam.  
  b. The nurse said that the bitter pill killed Sam.

- Experimental results: facilitation in both conditions
- Matches model predictions
- But: effect much weaker in non-coordinated context, in different measures
No coordination, same grammatical function:

(2) a. The young nurse and the bitter pill killed Sam.
   b. The young nurse said that the bitter pill killed Sam.

- Experimental results: facilitation in both conditions
- Matches model predictions
- But: effect much weaker in non-coordinated context, in different measures
There is a general mechanism at play, but context does play some role (size of the effect, type of coordinator: Knöferle and Crocker, 2007)

Phonological parallelism (Dubey et al., 2007): parallelism in size of coordinated constituents, as measured in syllables

Effect of frequency
- Stronger effect in high frequency settings
- Same pattern as seen in priming
- But does model make the right strength of prediction?
Do we need to hypothesise a special mechanism to model parallelism effect in coordination?

Argued by Frazier et al. (2000) and Apel et al. (2007)

However:
- Parallelism in coordination is a stronger instance of an effect which occurs everywhere
- General effect is enough to model experimental results
- Results from Frazier et al. (2000), and Knöferle and Crocker (2006) as well as our own experiments