Abstract

English and Japanese speakers have different word order biases in phenomena like heavy NP shift and lexical accessibility. These biases are expressed in sentence production, but are partially the result of language-specific learning. A connectionist model of syntax acquisition and sentence production was trained to produce English or Japanese sentences, and the resulting models were tested for their word ordering biases. The behavior of the models matched the human biases in each language and language-specific features of the input were found to be important in the development of these word ordering behaviors.

Word Order Biases in English and Japanese

- Heavy NP shift is a tendency for speakers to place long “heavy” noun phrases in non-canonical sentence positions.
- English speakers tend to place long noun phrases later in sentences (Stallings, MacDonald, & O’Seaghdha, 1998), while Japanese speakers tend to place these long phrases earlier in sentences (Yamashita & Chang, 2001).
- Accounts of heavy NP shift in these two languages (Hawkins, 2004) are not well integrated with theories of sentence production.
- Another word order bias is the tendency to place animate element earlier in sentences.
- Bransen, et al (2006) have found that animacy influences the position of syntactic functions in languages that allow free ordering of noun phrases such as Japanese.
- Both found that animacy does not influence the order of elements in noun conjuncts.
- Since sentence production theories make a rigid distinction between syntactic function assignment and position of words, it is hard for these theories to explain this cross-linguistic variability.
- Here, we examine whether these phenomena can be explained within a unified account of syntax acquisition and sentence production (Chang, Dell, & Bock, 2006).

A Connectionist Model of Syntax Acquisition and Sentence Production

- Chang, Dell, & Bock (2006) provided a connectionist model of syntax acquisition and sentence production that learns abstract syntactic representation in order to map between meaning messages and word sequences.
- Dual-path Architecture: sequencing system that learns syntactic representations, meaning system that learns lexical-concept links.
- Model accounts for a wide range of data (structural priming, syntax acquisition)
- Model trained on message-sentence pairs generated from input grammar.

Japanese and English Input for the Model

- The message encoded the concept-role bindings in the event as well as event-semantic information like tense and aspect. The message was paired with an English and Japanese sentence.
- Below is example dative.

```
MESSAGE:  MESSAGE:  
0A=SHOW 0X=BOY,DEF 0Y=ORANGE,DEF 0Z=DOG,DEF 0E=PRESENT,SIMPLE,AA,XX,ZZ,YY 0E=PRESENT,SIMPLE,AA,XX,ZZ,YY
```

English: The boy gave the orange to the dog.
Japanese: 犬にオレンジを与えた男の子。

- Japanese father
- English: The father gave

- Unlike previous models, the present model can handle messages with two propositions.

```
MESSAGE:  MESSAGE:  
0A=SHOW 0X=BOY,DEF 0Y=ORANGE,DEF 0Z=DOG,DEF 0E=PRESENT,SIMPLE,AA,XX,ZZ,YY 0E=PRESENT,SIMPLE,AA,XX,ZZ,YY
```

English: The boy gave the dog an orange.
Japanese: 犬にオレンジを与えた男の子。

- Differences between English and Japanese:
  - Head before relative clause
  - Head after relative clause
  - Pronouns: Argument omission
  - Alternations involve changes in syntactic functions
  - Alternations involve scrambling of order of arguments

Conclusion

A learning-based account of word order biases in English and Japanese sentence production

Franklin Chang

NTT Communication Sciences Laboratories, NTT Corp., Kyoto, Japan
chang@cslab.kecl.ntt.co.jp

Testing Heavy NP shift

```
English:  English:  
A girl went to the toy that rock.
Japanese:  a girl went to the toy that rock.
```

- The dependent measure was the number of Recipient before Patient structures. A mixed effect regression was performed with weight (long recipient, long patient) as a treatment factor on the data from the “adult” models (after 100000 patterns of training) for each language separately. The English model preferred to use the Recipient before Patient order when Patient were long (short-before-long bias, t = 7.98, p < 0.001), while the Japanese model preferred this order when the Recipient was long (long-before-short bias, t = 26.17, p < 0.001). Model has cross-linguistic word biases that are similar to English and Japanese speakers.

What causes heavy NP shift in the model?

- Several input factors were found to be important for heavy NP shift.
  - Ratio of simple main clauses to embedded clauses
  - Replacement of noun phrases with pronouns (English) or omission (Japanese)
  - Created four types of models by crossing ratio (High/Ratio/LowRatio) with pronoun status (WithPronouns/NoPronouns).
- Top figure: Found that mainembedded ratio was important for English (HighRatio) creates an expectation for short arguments in the postverbal positions, but not Japanese. This difference is due to the position of the shift in each language.
- Bottom figure: Found that the requirement that all arguments be expressed in Japanese (NoPronouns) caused all differences in weight to be minimized, which suggests that ambiguity in early sentence positions is important for Japanese heavy NP shift.

What causes heavy NP shift in the model?

- To examine animacy order, we created transitive and conjunct message-sentence whose elements varied in animacy.
- The model was tested with these elements and the percentage of animate before inanimate was recorded (test set had equal numbers of both orders)
- Top half shows that there is an animacy order preference in both languages for transitives, but not for conjuncts (t = 3.58, p < 0.001).
- Bottom two panels (Animacy Equated) shows that this bias is reduced when the input is completely balanced for animacy (unlike normal human input). Animacy order preferences in the model are learned from the distributional properties of the input.

Language-specific word order biases can be explained within a sentence production system that learns its representations from the input.

- In contrast to unified cross-linguistic accounts of heavy NP shift (Hawkins, 2004), the model suggests that the shift might be due to different language-specific factors in each language (structural frequency, argument omission).
- Rather than treating animacy order as a feature of a universal sentence production architecture, the model suggests that it is learned, and that is why it is expressed differently in different languages (syntactic functions, scrambling).

Accuracy of English and Japanese models over development

- Twenty training sets were created (each 40000 message-sentence pairs) and used to create twenty model subjects. Model was tested on 400 test sentences. Accuracy in terms of grammaticality and in terms of appropriate message was accessed. Although the languages differ greatly, the model can learn both to similar levels of accuracy.

References


Presented at the 21st Annual CUNY Sentence Processing Conference, Chapel Hill, 2008