Overview of PP-attachment ambiguities

Tyler J. Peckenpaugh

2019-05-30

Contents

Outline of the dissertation so far

What follows in this section is a brief overview of what is so far included in the dissertation.

Introduction

- What a garden path is, and how (the garden-path/Construal theory of) parsing works.[1]
- The PP-attachment ambiguity this paper is concerned with
- The intuitive observation that interrogative versions of these PP-attachment garden paths are less difficult to parse or recover from.

Background

- Prosody can effect parsing

[1] I think there's a fair chance this needs expanding. What do you see as being the necessary scope? Should I go through the whole development of the Frazier line of parsing theory?

    - Kjelgaard and Speer (1999),
- Prosody in silent reading
    - Fodor's (2002) Implicit Prosody Hypothesis[2]
- The prosody of questions vs. declaratives
    - Hedberg, Sosa, and Görgülü (2017)
- The details of how the parse of PP-attachment ambiguities leads to a garden path.
- Predictions

Methodology

- Recruitment
- Location
- Equipment and software
- Procedure
- Materials
- Groups of participants and versions of experiment
- IRT measurement
- Prosodic judgements

Results

- Data treatment
- Prosodic judgements
- IRT
- Delay comparison
- Demographic data and self-reported reading habits

Discussion

- Are the hypothesis supported?
- Behavioral correlate?
- Explaining the intuition
- Confounds
- Areas for further study

Conclusion

- Behavioral correlate of the intuition still might exist in IRT, but it remains to be fully supported (more data). Other possibilities exist (eye-tracking, ERP).

[2] Is this necessary? I supposed it's not...

- An explanation for the intuition still in the air, but the data seem to lean towards a non-prosodic account. Prosody-controlled embedded question work is the best next step.
- Other interesting findings:
  - Interrogativity has a robust impact on IRT.
  - Garden path condition has a robust impact on prosodic pattern.

Appendices

A. Experimental Items

B. Filler items

C. Instructions to participants

D. Instructions to research assistant on providing prosodic judgements

Specific requests for feedback

This section outlines some specific questions I have.

Predictions

These are the hypothesis I wrote before running the study, but they are not all really answered by the data I got. Is it dishonest to omit or revise them for this paper?

Hypothesis 1 High attachment of PP2 is prosodically marked by a prosodic break between PP1 and and PP2.

Hypothesis 2 A first reading of a GP sentence will exhibit less natural prosody (more hesitation at and after the disambiguating region) than:

- A first reading of a non-GP sentence.
- A second reading of a GP sentence.

Hypothesis 3 A first reading of a garden-path sentence will more often be produced with prosodic structure that represents an implausible or ungrammatical parse of the string (low attachment of PP2), whereas a previewed reading sentence will more often be pronounced with the prosodic structure that represents the intended parse (high attachment of PP2).

Hypothesis 4 A first reading of a declarative GP sentence will exhibit less natural prosody (more hesitation at and after the disambiguating region) and be more likely to be produced with prosodic structure that represents an implausible or ungrammatical parse of the string than a cold reading of an interrogative GP sentence.

Reliability of prosodic judgements

A second trained linguist repeated the task over 128 recordings selected from 8 participants (two from each group, one per ordering). Even number experimental items were used from 4 participants, and odd numbered from the other 4. There were 8 recordings missing from the 128 selected, so the reliability task resulted in judgements over 120 recordings. The first informat also blindly re-rated those 120, with the recording name obscured and instructions not to revisit her original ratings. Reliability scores (percent of recordings agreed upon) are reported in table 1.

Table 1: Inter and intrarater agreement with Cohen's Kappa

	Breaks			Break strength		Struggle		
	V	OBJ	PP1	STRONGEST	WEAKEST	STRUGGLED	START REGION	FINAL RISE
Interrater	50.0% K = 0.03 (z = 0.63)	65.0% K = 0.17** (z = 2.61)	78.3% K = 0.09 . (z = 1.86)	54.2% K = 0.25*** (z = 3.99)	23.3% K = 0.00 (z = 0.06)	85.0% K = 0.43*** (z = 5.23)	80.0% K = 0.27*** (z = 4.80)	95.8% K = 0.90*** (z = 9.88)
Intrarater	94.2% K = 0.34*** (z = 4.22)	77.5% K = 0.52*** (z = 5.73)	85.0% K = 0.52*** (z = 5.82)	72.5% K = 0.44*** (z = 5.70)	61.7% K = 0.38*** (z = 6.08)	92.5% K = 0.60*** (z = 6.62)	92.5% K = 0.58*** (z = 7.57)	95.8% K = 0.90*** (z = 9.95)

Note:
*** p < 0.001; ** p < 0.01; * p < 0.05, . p < 0.1

The lower intrarater agreement for relative break strength was likely a result of a methodological issue: it was possible to report the same pattern, e.g. a pattern where a PP1 break is stronger than an OBJ break, by either giving the response "PP1" for strongest break, and "OBJ" for weakest break; or, "PP1" for strongest and "NONE" for weakest; or, "NONE" for strongest and "OBJ" for weakest. While the instructions to the rater requested a particular method (avoid using "NONE" in these instances across the board), it's likely that inconsistencies occurred for these cases.

The same inconsistencies would have hurt interrater agreement for strongest/weakest also. A further contributing issue for interrater agreement of those two judgements stems from the poor agreement on the presence of the verb break. When the raters don't agree about the presence of a break, that disagreement is magnified for the judgement of the relative strength of breaks.

So, what am I to do about this spotty reliability?

Martin Chodorow on Cohen's K:

> Kappa measures the agreement over and above chance agreement. Consider a case where we have 2 raters, A and B, and two classification categories (Yes, No). If each rater is saying "Yes" 90% of the time and "No" 10% of the time, then we would expect them to agree on any given judgment 82% of the time, even if their judgments were independent. The reason for this is that if each source is generating Yes with probability of .90, then the probability that the Yes judgments will coincide by chance is .90 x .90 = .81, and if they are generating No with probability of .10, then the probability that those No judgments will coincide is .10 x .10 = .01 (so, .81 +.01 = .82). Let's call the probability of this kind of chance agreement Pc and the actual observed proportion of agreement Pa. The kappa formula is K = (Pa - Pc)/(1 - Pc). If, for my hypothetical example, the observed agreement were .91, then K = (.91 - .82)/(1 - .82) = .09/.18 = .50. When one of

the two classification categories has a high occurrence rate for both raters, it is difficult to get a very high kappa value. The table below shows the interpretations of kappa that are commonly used by researchers.

K	Interpretation
< 0	Poor agreement
0.00–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

PP1 and object breaks and their relative prominence

Table 2 presents data that incorporates the rater's judgement of the relative prominence of the breaks. The > symbol indicates that the rater found the break on the left of that symbol to be stronger, or more prominent, than the break on the left. When no symbol is shown between the two breaks, the rather found them to be of equal prominence.

Table 2: Reading 1 prosodic pattern type by condition

	D -GP	D +GP	Q -GP	Q +GP
OBJ	31.1	0.8	30.6	2.5
OBJ > PP1	32.0	23.0	29.8	14.2
OBJ PP1	5.7	3.3	1.7	4.2
PP1	14.8	26.2	24.8	25.8
PP1 > OBJ	16.4	46.7	13.2	53.3

Another way of looking at these same data is to think of a recording as being PP1-dominant (i.e., PP1 is the only or the strongest break), OBJ-dominate, or neither. This categorization is useful because it creates binary outcomes that can be subjected to logistic regression analyses. The frequency of these patterns is reported in table 3.

> Should "dominance" supplant or supplement the reporting of the simpler "PP1, OBJ, or both" reporting?

Table 3: Break dominance by condition, Reading 2 only

	OBJ		PP1	
	Not dominant	Dominant	Not dominant	Dominant
D -GP	37.4	62.6	69.1	30.9
D +GP	76.2	23.8	27.0	73.0
Q -GP	41.1	58.9	62.9	37.1
Q +GP	83.3	16.7	20.8	79.2

> This is perhaps a question for Martin, but: what should be made of the significance of interrogativity and GP:Q interaction?

Table 4: Logistic regression models of prosody, Reading 2 only

	OBJ	PP1 Dominance	OBJ Dominance
Intercept	2.196 ***	-0.947 **	0.595 *
	(0.413)	(0.288)	(0.258)
GP	-0.885 *	2.113 ***	-1.922 ***
	(0.355)	(0.317)	(0.311)
Interrogativity	-0.877 *	0.313	-0.174
	(0.352)	(0.292)	(0.279)
GP:Q Interaction	0.965 *	0.087	-0.324
	(0.481)	(0.435)	(0.440)
N	489	489	489
logLik	-240.448	-279.219	-278.314
AIC	492.896	570.438	568.628

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Reading habits and demographics

Some demographic information reported here is obtained from data in the QC recruitment system and other data are from a short hand-written survey adminstered after the experimental procedure. The survey questions were:

1. How often do you read magazines or newspapers for fun (i.e. not for school)?

   At least once a day At least once a week
   At least once a month
   Less than once a month

2. How often do you read fiction or non-fiction books for fun (i.e. not for school)?

   At least once a day At least once a week
   At least once a month
   Less than once a month

3. Do you speak any languages other than English natively and/or regularly?

   Yes No
   If yes, what language(s)?

4. When you read for fun, do you most often read in English?

   Yes, I typically read in English
   No, I typically read in another language
   Neither is true

5. How hard was it to read the sentences the way you wanted to?

   Not at all hard Somewhat hard
   Very hard
   Impossible

6. What do you think this study was about?

7. Do you have any suggestions for how your experience might have been improved?

8. Any other suggestions or comments?

For questions (1), (2), and (5), the responses were coded on a 1-4 scale during data entry, with the topmost answer being 4 and the bottom being 1. Question (3) was coded as 1 for "No" and 0 for any other answer. Table 6 shows the responses that are appropriate for data analysis.

What if any of this do you think would be interesting to analyze? My thought is to see if reading habit has an impact on IRT and/or break dominance. I looked briefly at some of this, and it's interesting to note that `semester` seemed to have an effect on IRT.

Table 5: Survey response to data input map

field	description
ID	Participant
Gender	Gender
lightReading	Survey question 1
bookReading	Survey question 2
Monolingual	Survey question 3
ease	Survey question 5
Semester	Semester the data were collected
DOTW	The day of the way data were collected
timeslot	The time of day the data were collected (24hr)
DATE	The date the data were collected

Table 6: Survey responses

ID	Gender	lightReading	bookReading	Monolingual	ease	Semester	DOTW	timeslot	DATE
1	F	4	1	0	3	Summer	FRIDAY	1300	6/22
2	F	2	3	0	3	Summer	MONDAY	1100	6/25
3	F	3	4	0	4	Summer	MONDAY	1200	6/25
4	F	3	2	0	3	Summer	MONDAY	1300	6/25
5	F	4	4	1	3	Summer	MONDAY	1400	6/25
6	F	2	1	0	3	Summer	MONDAY	1500	6/25
7	M	3	1	1	3	Summer	TUESDAY	1100	6/26
9	F	1	2	1	3	Summer	MONDAY	1200	7/2
10	F	4	1	1	3	Summer	THURSDAY	1100	7/5
11	M	4	4	0	4	Summer	THURSDAY	1500	7/5
12	F	1	2	1	4	Summer	FRIDAY	1200	7/6
13	F	4	1	1	3	Summer	MONDAY	1600	7/9
14	M	3	3	0	4	Summer	WEDNESDAY	1300	7/11
15	F	4	3	1	4	Summer	WEDNESDAY	1600	7/11
16	F	1	1	1	3	Summer	WEDNESDAY	1700	7/11
17	F	3	1	1	4	Summer	THURSDAY	1700	7/11
19	F	2	2	0	4	Summer	MONDAY	1500	7/16
20	F	3	1	1	4	Summer	WEDNESDAY	1500	7/25
21	F	1	1	0	4	Summer	WEDNESDAY	1600	7/25
22	F	3	3	0	4	Summer	WEDNESDAY	1700	7/25
201	F	3	2	0	3	Fall	WEDNESDAY	1200	11/14
202	M	1	4	0	2	Fall	WEDNESDAY	1500	11/14
203	F	1	3	0	3	Fall	WEDNESDAY	1400	11/14
204	F	4	1	1	3	Fall	WEDNESDAY	1700	11/14
205	F	2	1	1	3	Fall	TUESDAY	1300	11/20
206	F	3	1	1	4	Fall	TUESDAY	1400	11/20
207	M	3	1	1	3	Fall	TUESDAY	1500	11/20
208	F	1	1	1	3	Fall	TUESDAY	1600	11/20
209	F	3	2	1	4	Fall	WEDNESDAY	1200	11/21
210	F	3	1	1	4	Fall	WEDNESDAY	1400	11/21
212	M	4	1	1	3	Fall	WEDNESDAY	1600	11/21
214	F	1	3	0	4	Fall	WEDNESDAY	145	11/27
215	M	3	1	1	3	Fall	WEDNESDAY	1245	11/27