Analyzing Human and Machine Performance In Resolving Ambiguous Spoken Sentences

Hussein Ghaly[1] and Michael I Mandel[1,2]
[1] City University of New York, Graduate Center, Linguistics Program
[2] City University of New York, Graduate Center, Computer Science Program
{hghaly,mmandel}@gc.cuny.edu

Abstract

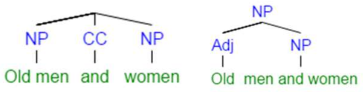
Written sentences can be more ambiguous than spoken sentences. We investigate this difference for two different types of ambiguity: prepositional phrase (PP) attachment and sentences where the addition of commas changes the meaning. We recorded a native English speaker saying several of each type of sentence both with and without disambiguating contextual information. These sentences were then presented either as text or audio and either with or without context to subjects who were asked to select the proper interpretation of the sentence. Results suggest that comma-ambiguous sentences are easier to disambiguate than PP-attachment-ambiguous sentences, possibly due to the presence of clear prosodic boundaries, namely silent pauses. Subject performance for sentences with PP-attachment ambiguity without context was 52% for text only while it was 72.4% for audio only, suggesting that audio has more disambiguating information than text. Using an analysis of acoustic features of two PP-attachment sentences, a simple classifier was implemented to resolve the PP-attachment ambiguity being early or late closure with a mean accuracy of 80%.

1 Introduction

There are different kinds of ambiguities in sentence construction, which can be challenging for sentence processing, both in speech and in text. Such ambiguities include structural ambiguities where there can be multiple parse trees for the same sentence. This includes coordination scope ambiguity, such as:

old men and women

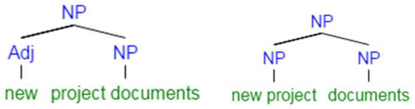
which can be parsed as either of the following trees with different meanings:



Another example is noun phrase ambiguity, such as:

new project documents

which can be parsed as either of the following trees, again with different meanings:



In speech, prosody has been shown to resolve certain ambiguities when the speaker is able to encode this information (Snedeker and Trueswell, 2003). In order to ensure that the speaker is able to do so, listening tests sometimes engage professional speakers, such as radio announcers, to read the sentence for maximum clarity (Snedeker and Trueswell, 2003).

In particular, Lehiste et al. (1976) found that the duration of words can resolve certain ambiguities reliably, specifically that syntactic boundaries can be perceived by listeners if the duration of the interstress interval at a boundary is increased. Price et al. (1991) found that some, but not all, ambiguities can be resolved on the basis of prosodic differences, where the

Proceedings of the First Workshop on Speech-Centric Natural Language Processing, pages 18–26
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

disambiguation is related more to the presence of boundaries and to some extent the prominence of certain words. However, when it comes to spontaneous everyday speech, especially by untrained speakers, Tree et al. (2000) found that although listeners can use prosody to resolve ambiguities, contextual information tends to overwhelm it when present. Krajalic and Brennan (2005) point out that results prior to their own study provide mixed evidence for whether speakers spontaneously and reliably produce prosodic cues that resolve syntactic ambiguities.

In text, punctuation can sometimes disambiguate the desired meaning. For example, the sentence:

1: A woman without her man is nothing

can mean:

1a: A woman, without her man, is nothing.

1b: A woman, without her, man is nothing.

The insertion of commas changes the meaning of the sentence so that it is not ambiguous when it is read. When each version is spoken, speakers also may encode cues to guide the listeners to the intended meaning. Typical automatic speech recognition output does not include punctuation, leading to transcripts that are ambiguous in this regard, even when the original speech might not be. One solution to this problem is to integrate a separate system for predicting punctuation from speech. For example, this has been done using neural network giving weights to different prosodic cues, where it was possible to predict 54% of the commas (Levy et al., 2012). Other methods include punctuation generation from prosodic cues to improve ASR output (Kim and Woodland, 2001). This is part of recovering the "structural meta-data" from speech, which also includes disfluencies and other sentence boundaries (Liu et al, 2006).

One of the most important ambiguities in both speech and text is prepositional phrase attachment (PP-attachment) ambiguity. A famous examples of this ambiguity is:

2: I saw the boy with the telescope.

In this case, no punctuation can help to resolve this structural ambiguity of whether the speaker or the boy had the telescope:

2a: I saw the boy [with the telescope]

2b: I saw [the boy with the telescope]

Snedeker and Trueswell (2003) have shown that this kind of ambiguity can be resolved by prosody in spoken sentences, cuing the different interpretations by the duration of the preposition itself (in this case: "with"), as well as the duration of the following phrase (in this case: "the telescope").

Because prosodic cues, when encoded by the speaker, can help guide the parsing of a structurally ambiguous sentence, we here explicitly compare the abilities of human listeners to disambiguate sentences in both written and spoken form, while starting to build a machine learning system that can perform the same task at least as well.

2 Hypothesis

The main hypothesis in this research is that when there is ambiguity in any sentence and the speaker is aware of the correct reading, they may convey their knowledge of the correct reading using certain prosodic cues. As Snedeker and Trueswell (2003) put it: "informative prosodic cues depend upon speaker's knowledge of the situation: speakers provide prosodic cues when needed; listeners use these prosodic cues when present."

Therefore, for sentences with comma ambiguity, given the correct punctuation, we can expect speakers to encode prosodic cues in their speech accordingly, and we can expect listeners to process these cues in their understanding of the sentence. For sentences with PP-attachment ambiguity, given a preceding disambiguating sentence, speakers may encode prosodic cues to indicate the intended meaning.

3 Goal

The ultimate goal of this research is to use prosody to improve parsing of ambiguous spoken sentences, allowing extracting information from speech that is not available from text only. This involves analyzing human disambiguation behavior for scripted sentences while building a machine learning system to automatically perform this disambiguation.

4 Data

Two types of sentences were investigated: sentences with comma ambiguities and sentences with PP-attachment ambiguity. We constructed 12 pairs of sentences with comma ambiguity and 14 pairs of sentences with PP-attachment ambiguity, as shown in the appendix.

4.1 Comma-ambiguous sentences

An example of a pair of comma-ambiguous sentences is:

3a: John, said Mary, was the nicest person at the party.

3b: John said Mary was the nicest person at the party.

These sentences are presented individually to the subject along with the question:

Who was said to be the nicest person at the party?
A: John
B: Mary

The correct answer for sentence 3a is A and for 3b is B.

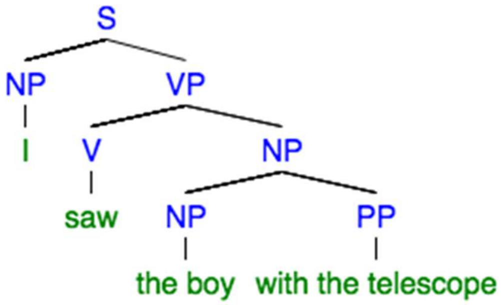
4.2 PP-attachment sentences

An example of a pair of PP-attachment ambiguous sentences is:

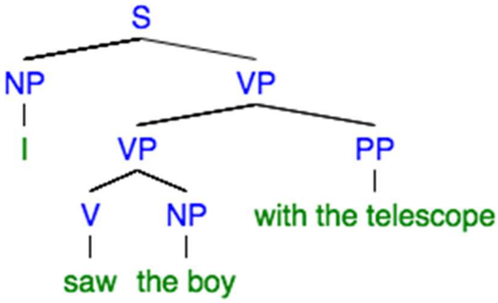
4a: *One of the boys got a telescope.* I saw the boy with the telescope.

4b:- *I have a new telescope.* I saw the boy with the telescope.

The initial italic sentence guides the speaker to the intended reading and in different experimental conditions were included or not included in the presentations to listening or reading subjects to measure their informativeness. The correct parse of sentence 4a exhibits "late closure":



The correct parse of sentence 4b exhibits early closure:



These sentences are presented individually to the subject along with the question:

Who has the telescope?
A: The boy
B: The speaker

The correct answer for sentence 4a is A and for 4b is B.

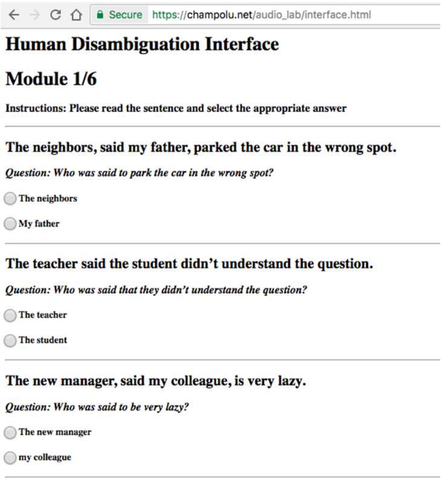
5 Method

5.1 Speech Data Collection

A native speaker of English recorded the complete list of 26 unique sentences, through a custom web interface implemented using Javascript and Python CGI. Each sentence was repeated five times and the 130 sentence instances were randomized before presentation to the speaker. PP-attachment ambiguous sentences were presented to the speaker with preceding context sentences, as in 4a and 4b. For the below experiments, all of the sentences with their text and audio are presented to the listeners.

5.2 Listener interface

Listener responses were also collected via another custom web interface. An example interface page is shown below:



Human Disambiguation Interface

Module 1/6

Instructions: Please read the sentence and select the appropriate answer

The neighbors, said my father, parked the car in the wrong spot.

Question: Who was said to park the car in the wrong spot?

○ The neighbors

○ My father

The teacher said the student didn't understand the question.

Question: Who was said that they didn't understand the question?

○ The teacher

○ The student

The new manager, said my colleague, is very lazy.

Question: Who was said to be very lazy?

○ The new manager

○ my colleague

5.3 Listener tasks

Sentences were presented to subjects either in written form or in recorded audio form. PP-attachment sentences were presented either with or without the preceding context sentence both for written and audio modalities. The tasks were presented in the following order, each one including a randomized ordering of all of the sentences:

1- Comma-ambiguity - Text
2- Comma-ambiguity - Audio
3- PP-attachment ambiguity with context - Text
4- PP-attachment ambiguity with context - Audio
5- PP-attachment ambiguity without context - Text
6- PP-attachment ambiguity without context - Audio

This order aims to familiarize the listeners gradually with the task by showing the text sentences first, which also serves as benchmark to detect any biases or confusion regarding the sentence itself. It then proceeds to the corresponding audio. The sequence follows a gradual increase of difficulty, saving for last the most difficult task: PP-attachment disambiguation without context in text and then audio.

6 Results

Four listeners participated in the study. Two of them were native English speakers. Their accuracy in identifying which of two possible meanings the speaker was cued is shown in the following table.

Ambiguity	Modality	Accuracy
Comma	Text	99.3%
Comma	Audio	94.7%
PP-attachment with context	Text	93.1%
PP-attachment with context	Audio	97.1%
PP-attachment without context	Text	52.0%
PP-attachment without context	Audio	74.4%

These results show that humans are quite good at interpreting comma-ambiguous sentences in both text and speech modalities. For PP-attachment, they also perform well for both modalities when the preceding context sentence is provided. Without the context sentence, they perform at chance for text, but much better than chance for speech, showing that there is, indeed, additional information present in the speech. Because performance is at ceiling for comma-ambiguity, we focus our subsequent analysis on the PP-attachment sentences.

The following table shows results for each of the PP-attachments sentences presented as speech without context. All productions of each version of each sentence are grouped together.

Sentence	Accuracy	N
1: I saw the boy with the telescope.	68.9%	29
2: I saw the man with the new glasses.	78.6%	28
3: San Jose cops kill a man with a knife.	89.3%	28
4: They discussed the mistakes in the second meeting.	70.9%	31
5: The lawyer contested the proceedings in the third hearing.	63.3%	31
6: He used the big wrench in the car.	82.1%	28
7: I waited for the man in the red car.	68.9%	29

In order to investigate the role of prosodic features in this disambiguation, we performed a preliminary semi-automatic analysis of the recordings of two of these sentences. A number of acoustic features were measured manually in Praat for all of the productions of both versions of two of the PP-attachment sentences, numbers

4 and 5. Following Levy et al (2012), we measured the following features:

- duration of the preposition utterance (in milliseconds)
- duration of the silent pause (if any) preceding the preposition (in milliseconds)
- duration of the noun phrase following the preposition (in milliseconds)
- Intensity of the preposition (in decibels)

By manually extracting features, we achieve an upper bound on the performance of an automatic feature extraction procedure.

In order to examine the minimum level of acoustic cues encoded by the speaker to see if it is still possible to extract meaningful patterns that can be used for automatic systems, we examine the sentences that listeners were unable to classify correctly.

As shown in the preceding table, one of the worst performing sentence for the PP-attachment disambiguation task from audio without context was:

4: They discussed the mistakes in the second meeting.

This sentence was correctly identified only 70.9% of the time, mostly being mistaken for early closure when in fact it was late closure, as shown in the detailed results in Appendix 2. This was not the case for this particular sentence for the audio with context or text with context.

The other sentence with most inaccurate disambiguation results (63.3% accuracy, evenly distributed between classes) was:

5: The lawyer contested the proceedings in the third hearing.

The following table shows the acoustic feature values averaged over the 20 productions of sentences 4 and 5. Note that both sentences use the same preposition and have the same number of words in the noun phrase following it.

	Late	Early
Preposition Duration (ms)	147	143
Preceding silent pauses (ms)	0	48
Intensity (dB)	57.84	56.37
Following NP duration (ms)	579	639.5

Using these data, we implemented a simple decision tree classifier to predict the closure type. Using 5-fold cross validation, the mean accuracy was 80%. The major node in the decision tree was the existence of a silent pause of smaller duration than 20 ms.

7 Conclusion

Although there has been much research in psychology regarding the perception of ambiguous sentences, more still needs to be done to model such sentences to facilitate integration with ASR systems, as well as question answering systems and natural language understanding.

The current research attempts to start developing this model. This is first done by quantifying human perception of certain ambiguous sentences, and analyzing these sentences acoustically to extract prosodic cues that can be used as features in a machine learning model for classifying sentences and deciding on their intended structure accordingly.

We found in our experiments that humans were able to disambiguate sentences with comma ambiguity at ceiling performance levels both as text and speech. For sentences with PP-attachment without context, human performance on text was close to chance at 52%, while for audio it was 74.4%, suggesting a richness of acoustic cues that can guide this ambiguation.

The machine learning model developed revealed the importance of the existence of a silent pause before the prepositional phrase as a major factor in determining the type of attachment. This, however, shouldn’t preclude the possible effects of other features and combinations thereof. For example, the average duration of the following NP was shorter for early closure than for late closure. These classifier results are preliminary given the very small size of the dataset.

Going forward, more speech samples need to be generated from multiple speakers. More listeners are needed to provide more certainty about the human ability to disambiguate. And these data can be analyzed in many more ways, both in terms of human perception and automatic classification.

As for extracting the acoustic features, a very important step is to use a forced alignment tool to measure the durations and starting and ending times for each word with greater accuracy and in a way that can be automated for a large number of speech files.

With more of both the human disambiguation data and acoustic data of the corresponding sentences, it will be possible to allow better parsing of ambiguous sentences from speech and the output of ASR systems.

8 Acknowledgements

We would like to thank Professors Janet Dean Fodor and Jason Bishop for their continuous support.

9 References

Kim, Ji-Hwan, and Philip C. Woodland. "The use of prosody in a combined system for punctuation generation and speech recognition." INTERSPEECH. 2001.

Kraljic, Tanya, and Susan E. Brennan. "Prosodic disambiguation of syntactic structure: For the speaker or for the addressee?." Cognitive psychology 50.2 (2005): 194-231.

Lehiste, Ilse, Joseph P. Olive, and Lynn A. Streeter. "Role of duration in disambiguating syntactically ambiguous sentences." The Journal of the Acoustical Society of America 60.5 (1976): 1199-1202.

Levy, Tal, Vered Silber-Varod, and Ami Moyal. "The effect of pitch, intensity and pause duration in punctuation detection." Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of. IEEE, 2012.

Liu, Yang, et al. "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies." IEEE Transactions on audio, speech, and language processing 14.5 (2006): 1526-1540.

Price, Patti J., et al. "The use of prosody in syntactic disambiguation." the Journal of the Acoustical Society of America 90.6 (1991): 2956-2970.

Snedeker, Jesse, and John Trueswell. "Using prosody to avoid ambiguity: Effects of speaker awareness and referential context." Journal of Memory and language 48.1 (2003): 103-130.

Tree, Jean E. Fox, and Paul JA Meijer. "Untrained speakers' use of prosody in syntactic disambiguation and listeners' interpretations." Psychological Research 63.1 (2000): 1-13.

Appendix 1 - List of Sentences

Sentence ID	Sentance	Type
1a	I have a new telescope. I saw the boy with the telescope.	late closure
1b	One of the boys got a telescope. I saw the boy with the telescope.	early closure
2a	She gave me new glasses. I saw the man with the new glasses.	late closure
2b	One of the men bought new glasses. I saw the man with the new glasses.	early closure
3a	Protests against knife-wielding cops. San Jose cops kill a man with a knife.	late closure
3b	Another man shot by the cops. San Jose cops kill a man with a knife.	early closure
4a	The project was full of mistakes. They discussed the mistakes in the second meeting.	late closure
4b	The second meeting was full of mistakes. They discussed the mistakes in the second meeting.	early closure
5a	The third hearing was full of problems. The lawyer contested the proceedings in the third hearing.	early closure
5b	The lawyer keeps complaining about the proceedings. The lawyer contested the proceedings in the third hearing.	late closure
6a	He bought a big wrench. He used the big wrench in the car.	late closure
6b	He was looking for any tool. He used the big wrench in the car.	early closure
7a	I rented a red car. I waited for the man in the red car.	late closure
7b	She told me he has a red car. I waited for the man in the red car.	early closure
8a	John, said Mary, was the nicest person at the party.	with commas
8b	John said Mary was the nicest person at the party.	without commas
9a	Adam, said Anna, was the smartest person in class.	with commas
9b	Adam said Anna was the smartest person in class.	without commas
10a	The teacher, said the student, didn’t understand the question.	with commas
10b	The teacher said the student didn’t understand the question.	without commas
11a	The neighbors, said my father, parked the car in the wrong spot.	with commas
11b	The neighbors said my father parked the car in the wrong spot.	without commas
12a	The new manager, said my colleague, is very lazy.	with commas
12b	The new manager said my colleague is very lazy.	without commas
13a	The author, said the journalist, didn’t address the main problem.	with commas
13b	The author said the journalist didn’t address the main problem.	without commas

Appendix 2- Detailed results by sentence for PP-attachment ambiguity

Ambiguous ?	Modality	Sentence ID	Mistake	Total
ambiguous	audio	1a	5	14
ambiguous	txt	1a	2	8
context	audio	1a	0	14
context	txt	1a	1	10
ambiguous	audio	1b	4	15
ambiguous	txt	1b	5	9
context	audio	1b	0	15
context	txt	1b	1	12
ambiguous	audio	2a	5	15
ambiguous	txt	2a	7	9
context	audio	2a	1	16
context	txt	2a	1	13
ambiguous	audio	2b	1	13
ambiguous	txt	2b	2	8
context	audio	2b	0	13
context	txt	2b	0	9
ambiguous	audio	3a	1	14
ambiguous	txt	3a	5	6
context	audio	3a	0	14
context	txt	3a	0	12
ambiguous	audio	3b	2	14
ambiguous	txt	3b	3	11
context	audio	3b	0	15
context	txt	3b	2	11
ambiguous	audio	4a	1	15
ambiguous	txt	4a	6	10
context	audio	4a	1	15
context	txt	4a	1	13
ambiguous	audio	4b	8	16
ambiguous	txt	4b	5	9
context	audio	4b	1	16
context	txt	4b	1	12
ambiguous	audio	5a	5	14
ambiguous	txt	5a	4	6
context	audio	5a	0	14
context	txt	5a	0	10
ambiguous	audio	5b	6	16
ambiguous	txt	5b	4	12
context	audio	5b	3	16
context	txt	5b	3	12
ambiguous	audio	6a	3	13
ambiguous	txt	6a	7	8
context	audio	6a	0	13
context	txt	6a	0	10
ambiguous	audio	6b	2	15
ambiguous	txt	6b	2	9
context	audio	6b	0	16
context	txt	6b	1	12
ambiguous	audio	7a	6	15
ambiguous	txt	7a	4	8
context	audio	7a	0	15
context	txt	7a	0	11
ambiguous	audio	7b	3	14
ambiguous	txt	7b	3	10
context	audio	7b	0	15
context	txt	7b	0	12

Appendix 3: Detailed feature values

Acoustic feature for productions of sentence 4:

File #	duration of preposition (ms)	preceding silence (ms)	following NP duration (ms)	Preposition Intensity (dB)	Closure Type
1	160	0	690	56.6	early
3	175	0	660	59.0	late
26	120	0	470	56.2	late
51	140	80	620	55.6	early
67	145	0	600	58.7	late
76	140	90	635	57.8	early
78	135	0	510	61.1	late
82	150	110	600	57.9	early
109	130	0	620	61.0	late
121	140	60	580	58.8	early

Acoustic features for productions of sentence 5:

File #	duration of preposition (ms)	preceding silence (ms)	following NP duration (ms)	Preposition Intensity (dB)	Closure Type
18	140	20	660	54.6	early
21	170	0	580	54.8	late
44	160	0	630	53.8	late
46	140	0	680	50.8	early
52	160	0	550	58.0	late
75	140	80	680	56.1	early
81	160	0	640	58.3	early
83	150	0	600	59.6	late
113	125	0	570	56.2	late
115	120	40	610	57.2	early