Pitch Range and Pause Duration as Markers of Discourse Hierarchy: Perception Experiments

Jörg Mayer, Ekaterina Jasinskaja & Ulrike Kölsch

Department of Linguistics
University of Potsdam, Potsdam, Germany

mayer@ling.uni-potsdam.de

Abstract
Discourse structure is reflected by a number of global prosodic parameters, like for example pause duration and pitch range. Discourse structure is also known to affect the accessibility/salience of antecedents of anaphoric expressions. Assuming these generalizations are correct, one can ask whether listeners use the information encoded in pauses and pitch range to resolve anaphoric references in ambiguous contexts. To examine this, we conducted a series of perception experiments with ambiguous discourses, where the pitch range of the sentences and the pause duration between sentences was manipulated. The results of our experiments corroborate our main research hypothesis that global prosodic parameters influence the resolution of anaphoric pronouns. The direction of the observed effect is clearly in accordance with the predictions of the existing theories of discourse anaphora and the current state of research on discourse prosody.

Index Terms: discourse structure, prosody, anaphora resolution.

1. Introduction

This paper presents the results of a study on the influence of prosody on the interpretation of anaphoric pronouns. Until now, empirical studies in this area have predominantly concentrated on local prosodic features of pronouns, such as pitch accent (cf. e.g. [1] and references therein), whereas the impact of global prosodic parameters of an utterance, such as pitch range or pause duration, was almost entirely ignored. At the same time, the current state of research on discourse anaphora on the one hand and global prosody on the other strongly suggests that a link must exist between these two (seemingly unrelated) phenomena. On the one hand, it is an established fact that discourse structure affects accessibility/salience of possible antecedents to anaphoric expressions [2, 3, 4, 5]. On the other hand, pitch range and pauses have been shown to signal the structure of a spoken monologue. With the present study we want to fill the gap in empirical research and check whether the expectation that global prosodic parameters affect anaphora resolution is correct.

1.1. Discourse structure and anaphoric accessibility

One of the factors that constrain the resolution of an anaphoric pronoun is the hierarchical discourse structure of the context in which the potential antecedent is mentioned. Roughly, if a new referent is introduced in a subordinated discourse unit (subtopic) it is no longer accessible for anaphoric reference after a shift from the subtopic back to the main topic. The following example illustrates this idea:

(1) a. Lena war glücklich nach dem Tennisturnier. 
   Lena was happy after the tennis tournament
   Lena was happy after the tennis tournament.

b. Die Silbermedaille war ein großer Erfolg.
   the silver medal was a great achievement
   The silver medal was a great achievement.

c. Die Trainerin gratulierte nach der Siegerehrung.
   the coach congratulated after the award ceremony
   The coach congratulated [her] after the award ceremony.

d. Für das nächste Turnier wünscht sie sich für the next tournament wishes she herself
   however the first place
   however the first place.

For the next tournament, however, she hopes for the first place.

The discourse in (1) allows for at least two possible structures. In both cases sentences (b) and (c) form a constituent that is connected to (a) by a subordinating discourse relation, e.g. Explanation: the segments (b) and (c) jointly present the cause of Lena’s happiness. For sentence (d), however, two attachments are possible. If it is attached higher up in the tree at the level of sentence (a) the predicted interpretation of the personal pronoun sie/she in the last sentence is Lena, since the referents introduced in the subordinated segment (b)-(c) are not accessible from (d)’s attachment site. On the other hand, if (d) is connected directly to the last segment the coach is an accessible antecedent. Since it is also the most recent one, the pronoun will preferably resolve to the coach.

The generalization illustrated above is captured in one form or another by almost any discourse theory, although the precise formulation may differ depending on the underlying assumptions about the discourse structure and the nature of anaphoric accessibility. One of the most well-known formulations is the Principle of the Right Frontier [3, 6, 7], which says that only the discourse units on the ”right frontier” of the discourse graph are accessible for any further operations, including search for anaphora antecedents, whereas the right frontier includes the immediate left sister of the current discourse unit plus all the dominating nodes, but not the subordinated nodes. A very similar constraint is proposed in [2], although it is formulated in more procedural terms. In this approach, discourse referents are organized in a stack of focus
bedded deep in the structure, so the pause between them will be (1d) relates directly to the preceding utterance, both nodes are em-
of sentence (1d) is expanded (see Fig. 1). By contrast, if sentence
pitch range of sentence (1c) is compressed, whereas pitch range
is likely to be associated with a longer pause and a pitch reset, i.e.
greatest structural break occurs immediately before it. This break
alizations. If the last sentence is attached higher in the tree the

1.2. Discourse prosody
Numerous studies have shown that the hierarchical structure of
spoken discourse is reflected by prosody. The two most important
and best researched prosodic means for structuring longer utter-
ances are pitch range and pause duration. Pitch range is a global
prosodic parameter of an intonational phrase and defines a subdi-
vision of the total range of fundamental frequency variation of a
given speaker. The pitch range can vary in width (e.g. expanded,
normal, compressed) and in position relative to the total range (e.g.
high, mid, low). It is the reference frame for local tonal events
like pitch accents and boundary tones. In general, most studies
agree that expanded pitch range correlates with the introduc-
tion of new discourse topics and sub-topics or with the beginning
of a paragraph; compressed pitch range, on the other hand, signals
the end of a paragraph or the closing of a (sub-) topic. These re-

result were already obtainable on the basis of a rather simplistic
pre-theoretical notion of discourse structure, equating the latter ei-
ther with the structure of a written text, i.e. the paragraph structure
[10, 11, 12] or with a discourse topic model adopted for the spe-
cific material of the study [13, 14, 15]. Other studies that based
their analyses on more elaborate, theoretically motivated hierar-
chical notions of discourse structure, such as Rhetorical Structure
Theory or Segmented Discourse Representation Theory, have also
shown that the width and position of the pitch range correlate sig-
nificantly with the depth of embedding of discourse units [16, 17].

Similar results are reported for the duration of silent pauses.
Pauses are longer before units introducing new discourse topics.
The shortest pauses appear between intonational phrases dealing
with the same topic [18, 15, 19].

These findings suggest that the two alternative discourse struc-
tures for the discourse in (1) should have different prosodic real-
izations. If the last sentence is attached higher in the tree the
greatest structural break occurs immediately before it. This break
is likely to be associated with a longer pause and a pitch reset, i.e.
pitch range of sentence (1c) is compressed, whereas pitch range
of sentence (1d) is expanded (see Fig. 1). By contrast, if sentence
(1d) relates directly to the preceding utterance, both nodes are em-
bedded deep in the structure, so the pause between them will be
shorter, and no pitch reset is expected (see Fig. 2). But given the
considerations of anaphoric accessibility discussed in the previ-
ous section, these prosodic contrasts should ultimately correlate
with the corresponding options of anaphoric pronoun interpreta-
tion: longer pause, pitch reset (high attachment in the discourse
structure) → pronoun sie/she is resolved to Lena; shorter pause,
no pitch reset (low attachment) → pronoun sie/she is resolved to
die Trainerin/the coach.

Although the predictions are obvious, it is still largely an open
question whether hearers actually use the information encoded
by pauses and pitch range to disambiguate anaphoric references,
which also bears on a more general theoretical issue, whether
global prosodic parameters contribute to the linguistic interpre-
tation of an utterance and should be represented in the grammar. In
general, perception studies on global prosody are relatively few
as compared to corresponding production studies, and even those
available concentrated mainly on more superficial aspects of per-
ception. It has been found for example that synthesized speech
with “paragraph intonation” sounds more natural than without it
[11], and that the hearers are able to identify discourse struc-
tural boundaries of different strength on the basis of prosodic cues
[15, 12]. As for the impact of global prosodic parameters on the
semantic/pragmatic interpretation of linguistic expressions, we are
aware of only one study by Silverman that addresses the issue [20,
chap. 6]. However, this study was based on very limited material
and the experiments were designed in such a way that the sub-
jects could easily guess the hypothesis under investigation. Thus
Silverman’s results, though encouraging, call for replication in a
methodologically more rigorous setting.

The purpose of the experiments presented in the following sec-
tions is thus to test the hypothesis that global prosodic features
contribute to the interpretation of linguistic expressions by disam-
biguating structurally ambiguous discourses. Among all the in-
terpretation phenomena sensitive to discourse structure, we confine
our attention to pronominal anaphora, which is a paradigmatic case
among various types of anaphora as well as a central phenomenon
of discourse interpretation.

2. Methods
2.1. Materials
2.1.1. Discourses
We constructed 28 discourses like the one in (1). All the discourses
comprised 4 sentences, where sentences 2 and 3 were connected
to sentence 1 with a subordinating discourse relation. Sentence 4
was constructed in such a way as to allow for two more or less
equally plausible interpretations: either as part of the embedded
sequence initiated by sentences 2 and 3 (low attachment), or as
related directly to sentence 1 (high attachment).

The critical discourse refers R1 and R2 (e.g. Lena and the
coach in (1)) were introduced in sentence 1 and sentence 3, respectively, and were not mentioned elsewhere until the target sentence 4, which contained an ambiguous pronoun that could refer to R1 or R2. R1 and R2 were realized by proper names or definite descriptions, always constituting the grammatical subject of the sentences and occurred in the pre-verbal position (the German Vorfeld). The ambiguous pronoun was the grammatical subject of sentence 4, too, but it was realized immediately after the finite verb (the first position of the Mittelfeld), whereas the preverbal position was occupied by a different constituent. The post-verbal position for the target pronoun was chosen because it is a prosodically weak position where the pronoun is most naturally realized without a pitch accent. Consistent deaccenting of the target pronoun in all the items was important, since accent placement on pronouns is known to affect anaphora resolution (see e.g. [1]).

Additionally, 36 fillers were constructed. The filler discourses also comprised 4 sentences but showed no ambiguity in discourse structure or pronoun resolution. Each discourse (experimental and filler) was completed with a final who?-question of the form in (2). In the experimental items, the question was derived from sentence 4, as to reveal the hearer’s interpretation of the target pronoun.

(2) Wer wünscht sich den ersten Platz?
Who hopes for the first place?

2.1.2. Experimental items

All materials were recorded in an anechoic chamber using high quality equipment. The sentences were read by one female speaker in randomized order (i.e. not in the context of the respective discourses), aiming at producing constant pitch range and intensity settings. We then adjusted the pitch range parameters (see 2.1.3) for each sentence and re-created the original discourses by concatenating the resynthesized sentences with specific pause durations (intervals of zero amplitude). Concerning the 28 experimental discourses, 2 versions of each discourse were created resulting in 56 experimental items.

Experiment 1 (pause and pitch range): In the low attachment version, pauses were set to standard duration (400 ms) between all sentences. The pitch range of sentences 2 and 3 was set to normal, and pitch range of sentence 4 was compressed. The high attachment version had a lengthened pause (800 ms) between sentence 3 and sentence 4 and standard pause durations between the other sentences. Pitch range was set to normal in sentence 2, compressed in sentence 3, and expanded in sentence 4, i.e. there was a pitch reset between sentences 3 and 4. In both versions, sentence 1 was always assigned expanded pitch range (cf. Figures 1 and 2).

Experiment 2 (pitch range only): All pauses in both versions were set to standard duration. High and low attachment versions differed only in the pitch range settings (like in Exp. 1).

Experiment 3 (pause only): The pitch range of all sentences in both versions was set to normal. High and low attachment versions differed only in pause settings (like in Exp. 1).

For the 36 fillers, pitch range and pause durations were set according to their discourse structure, which either matched one of the discourse structural patterns of the items, or exhibited a third pattern where the greatest structural break occurred between sentences 2 and 3.

The final question was spoken by a male speaker and was added to the sequence 1500 ms after the end of sentence 4 with the original unmanipulated question intonation.

2.1.3. Pitch range manipulation

Pitch range was defined as the range between the highest intonationally relevant high tone (HT) and the lowest relevant low tone (LT) within one phrase. HT and LT were labeled manually in the original recordings and corresponded usually to high or low tonal targets of pitch accents. For pitch range manipulations, 3 different ranges were defined: normal, compressed and expanded. We determined the normal pitch range of the female speaker as ranging from 150 Hz (baseline) to 270 Hz (topline). Compression and expansion ratios were calculated from radio news corpus. Therefore, the differences to normal range were relatively moderate and less distinctive compared to spontaneous speech:

<table>
<thead>
<tr>
<th></th>
<th>baseline</th>
<th>topline</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>150 Hz</td>
<td>270 Hz</td>
</tr>
<tr>
<td>expanded</td>
<td>150 Hz</td>
<td>310 Hz</td>
</tr>
<tr>
<td>compressed</td>
<td>150 Hz</td>
<td>250 Hz</td>
</tr>
</tbody>
</table>

The pitch contour of each sentence was shifted so that the LT was set to the target baseline and multiplicated so that the HT reached the target topline. Based on these manipulated contours, the sentences were resynthesized using PSOLA resynthesis techniques implemented in Praat [21], with virtually natural sound quality.

2.1.4. Expectations

Regarding Experiment 1, we expected that listening to the low attachment version with a short pause between sentence 3 and 4 and gradually decreasing pitch range over the whole sequence, hearers would preferably attach sentence 4 inside the subordinated constituent. This would be indicated by a preferred co-reference between the pronoun in sentence 4 and the second referent R2 introduced in sentence 3. Listening to the high attachment version with a long pause and pitch reset between sentence 3 and 4, the hearers would attach sentence 4 outside the subordinated constituent more frequently than in the first condition, which would be indicated by a more frequent resolution of the target pronoun to the referent R1 introduced in sentence 1. Experiments 2 and 3 were designed to test whether pitch range alone or pause structure alone were sufficient to trigger this effect.

2.2. Subjects and procedure

36 subjects participated in experiment 1, 29 in experiment 2, and 33 in experiment 3. All subjects were native German speakers. For all 3 experiments, two item sets were compiled from a total of 56 experimental items (28 x 2 versions) and 36 filler items. Each set contained only one version of the experimental items and all filler items, resulting in a total of 64 items per set in randomized order (14 experimental items/high attachment version + 14 experimental items/low attachment version + 36 fillers). Half of the subjects were given set A, the other half set B, so each subject heard each discourse exactly once. Subjects were asked to listen to the stories and answer the question at the end of each story orally. Neither written transcripts nor lists of possible answers were provided. The answers were immediately classified by the experimenter (high attachment, low attachment, indefinite).

3. Results

The participants’ responses were coded on a nominal scale with 1 for R1 (high attachment response) and –1 for R2 (low attachment response).
response). The data were aggregated within the experimental condition and two tailed paired t-tests were done for all participants (subject analysis: t1) and items (item analysis: t2).

3.1. Experiment 1

Besides a predominance of the low attachment response in both conditions which can be explained by the general preference for the nearest antecedent (referred to sentence 3), we found a significant difference between conditions (t1 [35]=2.992, p < 0.01) and t2 [27]=4.804, p < 0.001), i.e. the high attachment response was significantly more frequent in the high attachment prosody condition (39%) than in the low attachment prosody condition (28%; cf. Table 1).

3.2. Experiments 2 and 3

Again, we found a predominance of the low attachment response in both conditions of experiments 2 and 3. In contrast to experiment 1, however, no significant difference between conditions was found, neither in experiment 2 (pitch range only) nor in experiment 3 (pause structure only) (cf. Table 1).

4. Conclusion

The results of our experiments corroborate our main research hypothesis that global prosodic parameters such as pitch range and pause duration influence the resolution of anaphoric pronouns. But experiments 2 and 3 revealed apart from this, that only the combination of both prosodic parameters has enough strength in order to disambiguate structurally ambiguous discourses. The direction of the observed effect in experiment 1 is clearly in accordance with the predictions of the existing theories of discourse anaphora and the current state of research on discourse prosody. This result strongly suggests that, indeed, the relationship between global prosody and the choice of the antecedent of an anaphoric pronoun is mediated by the choice of attachment site of an utterance in the discourse structure. The current results are also consistent with those obtained by Silverman [20]. Our experiment can be viewed as a replication of Silverman’s result in a methodologically more rigorous setting. Furthermore, the current experiment complements Silverman’s study by testing the hypothesis on a different discourse structure sensitive phenomenon (pronominal anaphora) and in a different language (German). The two studies taken together present substantial evidence for a linguistically relevant interaction between global prosodic parameters and discourse structure.

5. References