Audiovisual Congruence and Pragmatic Focus Marking

Charlotte Wollermann , Bernhard Schröder , and Ulrich Schade

Institute of Communication Sciences, University of Bonn, Germany
German Linguistics, University of Duisburg-Essen, Germany
Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Germany
cwo@ifk.uni-bonn.de, bernhard.schroeder@uni-due.de, ulrich.schade@fkie.fraunhofer.de

Abstract
This paper presents an empirical study on the interplay between audio and visual information for pragmatic focus marking. Nine German speakers were instructed to read dialogues with embedded question-answer pairs and varied context regarding certainty and exhaustivity. Results show that H* accompanied by a raising of eyebrows or head occurs significantly more often for the context intended to advantage uncertainty and non-exhaustivity. Furthermore, when two noun phrases are coordinated, a higher number of audiovisually equivalent realizations occur for the context intended to advantage uncertainty and exhaustivity, whereas audiovisually incongruent cues occur more often for the context intended to advantage uncertainty and non-exhaustivity. We do not consider natives (intended to advantage exhaustivity) or included them

1.1. The concept of pragmatic focus
The term focus often refers to the intuition that pitch accent correlates with new information in utterances, whereas old information is deaccented, (at least) in West Germanic languages like English, German or Dutch [1]. Even though focus is a complex phenomenon and labelled differently [2], most theories agree that focus can be defined as the answer to an – explicitly or implicitly given – question [3]. For our study we use this concept of focus, the pragmatic focus. In [4, 5] it is assumed, that focus is associated with such a background question. If the latter is interpreted as a mention-all question, the precondition for an exhaustive interpretation is given.\footnote{We would like to thank Bernd Möbus and Kellyn Rein for helpful comments. Many thanks to Denis Arnold, Natascha Blotzki and Anne Tielke for helping us annotating the data and to our speakers.}

(1a) Who kissed Mary?
(1b) John kissed Mary.

If the hearer of (1b) concludes that John is the only person who kissed Mary, she interprets the answer exhaustively. Whether the hearer interprets the speaker’s answer exhaustively depends on the knowledge about the answer which is ascribed to the speaker by the hearer [6]. In our scenario, we assume that if the speaker signals (unc)ertainty she is (not) completely informed; we only test exhaustive interpretation which is epistemically strong in the terms of [6]. In [4, 5] it is argued that, in the context of an adequate question, pitch accent is highly correlated with focus and then an exhaustive interpretation should be facilitated. The empirical findings of [2] suggest that not only accent is relevant for focus interpretation, but also the hearer’s expectations, the sensitization for focus phenomena, contextual factors and structural factors. Similarly, [7] reported for written speech that the type of question and answer affects exhaustivity.

1.2. Audiovisual cues of uncertainty
The empirical work of [8, 9] reported that uncertainty is expressed and also perceived by rising intonation, delays, fillers and lexical cues. [10] found that smiles and “funny faces” are visual indicators of uncertainty. [11] suggested for English that fall-rise intonation contributes to a context-independent meaning of utterance interpretation conveying the speaker’s uncertainty, but for German it has been barely studied which prosodic cues express uncertainty on the pragmatic level.

1.3. Pitch and visual prosody
According to the metaphor of up and down [12] rise or fall of pitch is accompanied by rise or fall of gesture and facial expression. Empirical evidence for this assumption was found in [13], where a rising of eyebrows (especially the left) was accompanied by a rising of pitch. Further, the data of [14] suggested a correlation between head movement and pitch.

1.4. Assumption
We assume that if the speaker signals uncertainty – due to the lack of knowledge – she will convey it audiovisually. The hearer will use this information for decoding the utterance, thus a non-exhaustive interpretation should be preferred.

2. Related work
In [15] we tested the influence of intonation and micro context on the exhaustivity of answers. Our results suggested that the exhaustive reading was generally preferred, but falling intonation (intended to convey certainty and finality according to the biological codes [1]) for focus marking in combination with a question which is congruent to the answer was biased toward exhaustivity. In contrast, rising intonation (intended to convey uncertainty and finality) combined with a general question was biased toward non-exhaustivity. In a follow-up study [16] we varied intonation for the focus constituent and the sentence-final verb and also pauses. The macro context either excluded alternatives (intended to advantage exhaustivity) or included them...
We observed strong effects of the macro context on non-exhaustivity, but the impact of prosody was relatively weak. From these findings we derived a model of pragmatic focus interpretation [17]: micro and macro context is relevant for raising the hearer's expectations; these expectations and prosodic information influence focus interpretation. Therefore we regard it as necessary to test the impact of context on pragmatic focus realization.

3. Production study

3.1. Goal

The goal of our study is to test which audiovisual cues speakers use when uttering (non-)exhaustive answers and if there are correlations with audiovisual cues of (un)certainty and (non-)finality.

3.2. Material

The stimuli were based on the material used in [15, 16] and consisted of six question-answer pairs embedded in different dialogues. The scenario was a fictitious party where different groups of students acted differently. For every action, there was a question asking about the agent and an answer providing the information. The focus exponent in the answer was either one group of students or two groups. We refer to one group by focus sentence with one NP (noun phrase) and to two groups by focus sentence with (a coordination of) two NPs. The ratio was 50:50. Two variants of context were generated. i) The variant exh+ was characterized by contextual congruity. One student group was salient during the dialogue, a question followed which was congruent to the focus utterance (see 2a + 2c). Thus alternatives were excluded by the context (only the mathematicians and designers disturbed the neighbours by laughing loudly). Further, a sentence indicating certainty about the answer followed (see 2d). ii) The variant exh- was marked by contextual incongruity. One “competing” discourse entity was introduced at the beginning of the dialogue (usually the linguists laugh loudly). A general question followed (see 2b) and the respective students group as focus of the answer (see 2c) was different from the salient discourse entity, i.e., alternatives were included. A sentence indicating uncertainty about the answer followed (see 2e). We assumed that the variant exh+ should advantage the production of audiovisual cues of certainty and finality, whereas the variant exh- should advantage the realization of audiovisual cues of uncertainty and continuation.

(2a) Wer hat die Nachbarn durch lautes Lachen gestört? Who disturbed the neighbours by laughing loudly?
(2b) Was ist passiert? What happened?
(2c) Die Mathematiker und Designerinnen haben die Nachbarn durch lautes Lachen gestört. The mathematicians and designers disturbed the neighbours by laughing loudly.
(2d) Das waren die Einzigen, da bin ich mir sicher. I am certain that they were the only ones.
(2e) Könnte aber sein, dass die Linguisten auch mitgelaucht haben. It is possible that the linguists also laughed.

3.3. Procedure

Nine speakers (1 m, 8 f) were recorded, all of them students at Bonn University and native speakers of German. We used scripted dialogues since we gave top priority to the checkability and comparability of the data. Subjects were instructed to read out loud the six stimulus dialogues and three further filler dialogues, presented in a random order. For all speakers the dialogue partner was always the same person.

3.4. Audiovisual annotation

We annotated the NP(s) as focus constituent(s) in the answer. For the audio annotation we used GToBI [18]; for the visual information we used ELAN to annotate the presence of eyebrow and head movement. There were three coders for each modality.

Table 1 shows the accent types observed in our data and table 2 the visual cues.

3.5. Analysis of speakers and dialogues

In [19] we previously looked at the data by subsuming all speakers and all dialogues. We found for the realization of the focus sentence, that in the case of two coordinated NPs, the first NP was significantly more often marked by L+H* for the variant exh+, whereas H* was produced more often for exh-. Furthermore, we observed for the realization of the first NP raised eyebrows or head accompanying H* more often for exh- in a marginally significant way. Since our first analysis revealed that significant differences between audiovisual realization of pragmatic focus occurred exclusively for focus sentences with two NPs it is necessary to do a more fine-grained analysis of these data.

3.6. Analysis of dialogues with two NPs as focus constituents

We analysed the audiovisual realization of the focus constituent of 27 focus sentences (9 speakers x 3 dialogues) with respect to three questions. I) Do different audiovisual realizations occur for each of the two variants? II) Is there a difference between the audio and/or visual realization of the first and second NP? III) Do the same audio and/or visual patterns occur for realizing the first and second NP more often for the variant exh+? We assume that speakers use audio and/or visual congruity more often for exh+ since the context is marked by a higher degree of contextual congruity than for exh-, i.e., the congruity between (micro and macro) context and the focus of the answer.

Table 3 shows, for each of the three dialogues, the prosodic realizations for the first and second NP. For the statistical analysis Fisher’s exact test is used. With respect to question I we observe the following when we cluster the data for the first and second NP, but distinguish between the two variants (exh+: N=26, exh-: N=28). For the variant exh+ L+H* alone (L+H*, E-, H-) is the realization which appears most often with seven, the absence of all three cues (acc-, E-, H-) is observed six times. Furthermore, L+H* combined with raised eyebrows (L+H*, ER, H-), L+H* accompanied by head lowering (L+H*, E-, HL), L+H* in combination with raised eyebrows and head lowering (L+H*, ER, HL) and also raised eyebrows alone (acc-, ER, H-) is produced each time once. The following prosodic realizations are found each time once: H* alone (H*, H-, H-), L+H* combined with frowned eyebrows (L+H*, EF, H-), L+H* accompanied by head raising (L+H*, E-, HR), L* alone (L*, E-, H-) and head lowering alone (acc-, E-, HL). For the variant exh- L+H* alone (L+H*, E-, H-) is found most often with 10, and, similar to exh-, it is followed by absence of all three cues (acc-, E-, H-) with four. H* accompanied by raised eyebrows (H*, ER, H-) is produced three times; the same accent type combined
Table 1: Accent types

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>H*</td>
<td>peak accent</td>
</tr>
<tr>
<td>L*</td>
<td>low accent</td>
</tr>
<tr>
<td>L+H*</td>
<td>rising peak accent</td>
</tr>
<tr>
<td>L+H</td>
<td>low rising accent</td>
</tr>
<tr>
<td>ac-</td>
<td>no accent</td>
</tr>
</tbody>
</table>

Table 2: Visual cues

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER</td>
<td>eyebrows raised</td>
</tr>
<tr>
<td>EF</td>
<td>eyebrows frowned</td>
</tr>
<tr>
<td>E-</td>
<td>no eyebrow movement</td>
</tr>
<tr>
<td>H*</td>
<td>head raised</td>
</tr>
<tr>
<td>H*</td>
<td>head lowered</td>
</tr>
<tr>
<td>HS</td>
<td>head shaking</td>
</tr>
</tbody>
</table>

Table 3: Prosodic realizations for focus sentences with two coordinated NPs (for explanations see table 1 and 2); (2) marks that the realization is found twice; dark grey marks audiovisual congruity between all three cues, light grey marks audiovisual incongruity (at least one incongruent cue); bold marks realizations of H* combined with eyebrows or head raised

<table>
<thead>
<tr>
<th>Dialogue 1</th>
<th>Dialogue 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exh</strong></td>
<td><strong>Exh</strong></td>
</tr>
<tr>
<td>L+H*, E-H,</td>
<td>L+H*, E-H,</td>
</tr>
<tr>
<td>acc- E-, H-</td>
<td>acc- E-, H-</td>
</tr>
<tr>
<td>2nd NP</td>
<td>2nd NP</td>
</tr>
<tr>
<td>Congruity</td>
<td>Congruity</td>
</tr>
<tr>
<td>Deaccentuation</td>
<td>Deaccentuation</td>
</tr>
</tbody>
</table>

with raising of head (H*, E-, HR) and also H* alone (H*, E-, H-) each time twice. The following realizations are produced once: L+H* in combination with head shaking (L+H*, E-, HS), L+H* combined with head lowering (L+H*, E-, HL), L+H* accompanied by raised and then frowned eyebrows (L+H*, ER/F, H-), L+H* combined with raised eyebrows, but head lowering (L+H*, ER, HL), H* with frowned eyebrows and head raising (H*, E, HR), raised and then frowned eyebrows alone (acc-ER, E/H) and L+H* alone (L+H*, E-, H-). The analysis shows that H* accompanied by a raising of eyebrows (ER) or head (HR) occurs more often for the variant exh- than for exh+ (6:22 vs. 0:26 vs. p 0.02). This result is in line with our findings reported in [19].

Concerning question II, the following can be reported investigating the difference between the production of the first and second NP independent of the two intended variants: for dialogues 1 and 2 there is no significant difference between the numbers of marking of accent, of eyebrows, of head and also of accent combined with eyebrows and/or head (each time p 0.05). For dialogue 3 accentuation does not occur significantly more often for realizing the first NP, but raising of eyebrows or head is produced significantly more often for the first NP (5:4 vs. 0:9, p 0.03). The marking by accent combined with either eyebrows or head or is also observed more often for the first NP in a significant way (5:4 vs. 0:9, p 0.03). Taking all the data together (exh+ and exh-: each time N=27) (see fig. 1a) our data show that accentuation appears more often with the first NP than with the second NP in a marginally significant way (24:3 vs. 17:10, p 0.05). Eyebrow movement is produced significantly more often for the first NP (10:17 vs. 2:25, p 0.02), but this is not the case for head movement alone (8:19 vs. 5:22, p 0.53). Either eyebrow or head movement are found more often for the first NP in a significant way (17:10 vs. 6:22, p 0.003) and accentuation combined with eyebrows and/or head movement as well (24:3 vs. 8:19, p 1.746e-05).

Regarding question III, we find neither for dialogue 1 nor for dialogue 3 a significant difference regarding audio and/or visual congruity between the first and the second NP when comparing the two variants. The p-value for differences regarding the congruity of accent, of eyebrows, of head and also of accent combined with both eyebrows and head is each time p 0.05. For dialogue 2 our data reveal the occurrence of exactly the same prosodic pattern of all three cues, i.e., accent, eyebrows and head, significantly more often for the variant exh- than for exh+ (4:1 vs. 0:4, p 0.048). When we cluster the data by subsuming all three dialogues the following is observed (see fig. 1b): the two variants (exh+: N=13, exh-: N=14) do not differ significantly with respect to the congruity of accent (7:6 vs. 6:8, p 0.7), eyebrows (10:3 vs. 7:7, p 0.2) and head (11:2 vs. 11:3, p 1), but the same pattern of all three cues occurs more often for exh+ than for exh-. This difference is marginally significant (6:7 vs 2:12, p 0.1). In accordance, our data reveal for the variant exh- a marginally significant higher number of audiovisual incongruities, i.e., realizations with at least one incongruent cue (7:6 vs. 12:2, p 0.1). For exh+ in five cases (of incongruity) we observe deaccentuation for the second NP: deaccentuation of accent alone and also of head alone appears each time once and deaccentuation of both accent and eyebrows three times. For p exh- we find deaccentuation of accent and head once, deaccentuation of eyebrows alone four times, deaccentu-
4. Conclusion

We presented an investigation of the role of context for the audiovisual marking of pragmatic focus. Our data suggest differences between the dialogues for prosodic realizations. However, in general we find, independent of the context, significantly more often accentuation and/or eyebrow movement for producing the first NP than for the second (coordinated) NP. This stronger prosodic marking of the first NP might be explained by the fact that the speaker regards this constituent as particularly important and thus emphasizes it for the hearer. For the context intended to advantage non-exhaustivity and uncertainty, a significant occurrence of H* accompanied by a raising of eyebrows or head is observed. We interpret this as audiovisual manifestation of the biological codes according to [1] and also as evidence of the metaphor of up and down [12]. Furthermore, we observe that speakers use the same patterns of all three cues, i.e., accent, eyebrows and head, for realizing the two coordinated NPs generally more often for the context characterized by incongruent features, intended to advantage exhaustivity and certainty. In contrast for the context characterized by incongruent features, intended to advantage non-exhaustivity and uncertainty, a higher number of incongruent prosodic patterns for the coordinated NPs is found, in particular deaccentuation of eyebrows. Our results are in line with [20] who reported the realization of same accent types for list items of read-aloud lists. We infer that when context and focus utterance differ from each other, incongruity of audiovisual cues appears even though dialogues are read loud and not spontaneous. In our future work we will test the effect of different audiovisual realizations for pragmatic focus interpretation. For further research on the production side we consider it important to use spontaneous speech in order to test if audiovisual congruity occurs when speakers signal exhaustivity and use uncertainty as a paralinguistic expression of their own information state instead of eliciting it by a scripted context.

5. References