Perceiving focus

Krahmer, E.J.; Swerts, M.G.J.

Published in:
Topic and focus

Publication date:
2007

E. KRAHMER [(1)] AND M. SWERTS [(1,2)]

PERCEIVING FOCUS

(1) Tilburg University, Communication & Cognition
(2) Antwerp University, Center for Dutch Language and Speech

Abstract. This chapter presents a series of studies that investigate how listeners use pitch accents to determine the focus of an utterance. We first motivate why we believe that a perceptual approach is necessary to gain insight into the importance of pitch accents. We outline a methodological approach which consists of a particular procedure to elicit speech data and a specific perceptual task which amounts to detecting the main focus in an utterance (a dialogue reconstructing experiment). We first present results of experiments using Dutch speech-only stimuli, and then compare the results obtained this way with those for Italian data, given that this language has been claimed to be markedly different in terms of accent distribution. Then, we explore how functional pitch accents are in multimodal stimuli, using a Talking Head which also uses rapid eyebrow movements to signal focus. Results show that the relative importance of pitch accents for the perception of focus is dependent on the language system and on the communicative setting in which they occur.

Keywords: Prosody, focus, perception experiments, cross-linguistic comparison, multimodality

1. INTRODUCTION

Many linguists approach intonational matters from a purely speaker-oriented perspective[1]. For instance, in different studies, in as far as these are empirical in nature, evidence for particular tonal distinctions is often solely based on acoustic analyses of fundamental frequency (F0) traces. However, if one wants to gain full insight into how intonation 'functions', such an approach is arguably incomplete. That is, a prosodic feature, as any other linguistic feature, can only be said to be communicatively relevant if it is not only encoded in the speech signal by a speaker, but if it also has an impact on how an utterance is processed by a listener. In other words, claims about important intonational categories and their respective meanings are somewhat premature if they are not backed up with results that show that these are also relevant at the receiving end of the communication chain. Ideally, such an analysis should be more than an individual linguist's interpretation of a prosodic phenomenon.

[1] *This chapter presents an overview of our work on the perception of focus, a research topic that we have been involved with since 1998. The studies focusing on the dialogue reconstruction for Dutch and Italian are presented with more detail in Swerts et al. (2002). A preliminary version of the third, audiovisual study is described in Krahmer et al. (2002a). Thanks are due to our colleagues Cinzia Avesani, Zsófia Ruttkay and Wieger Wesselink for their help in carrying out these studies.*

T.H.E. Editor(s) (ed.), Book title, 1—6.

Unfortunately, one cannot simply take it for granted that all prosodic detail really matters to a listener. One obvious, but sometimes neglected, condition is that tonal variation clearly needs to be above a perceptual threshold to be functionally relevant. In that respect, it is striking to see that many researchers attach functional load to particular tonal distinctions, which, from a purely phonetic point of view, are only minimally separable or even highly overlapping in "tonal space". For instance, the difference between H* and L+H*, as defined in the ToBI framework, has been claimed to indicate semantically distinct categories such as rheme and theme (Steedman 2000) or new and contrastive information (Pierrehumbert & Hirschberg 1990). Yet these two intonational categories are often confused by labellers who are instructed to transcribe intonation (e.g. Pitrelli et al. 1994), even to the extent that some investigators simply give up on the distinction. In comparison, many vowel systems of the world obey a contrast principle, which states that any two vowels need to be optimally distinct in order to be appropriately applicable in speech communication (the idea of vowel dispersion, see e.g., ten Bosch 1991). Also, linguistic systems are highly redundant in that speakers have various strategies at their disposal to signal particular meanings. Since tonal markers of semantic events often covary with morpho-syntactic, lexical or other prosodic cues, it is theoretically possible that their communicative function is 'overruled' by that of other resources, or by the situational or linguistic context in which they occur.

In this chapter, we argue that controlled perceptual studies allow us to investigate the communicative importance of intonational features. Rather than concentrating on subtle differences between intonational categories, we will illustrate this viewpoint with a series of studies on the cue value of pitch accents. In languages such as Dutch and English, the distribution of accents has been claimed to be exploited as a means to distinguish important bits of information in an utterance from unimportant ones. That is, in such languages, pitch accents serve as a linguistic strategy to put 'new' or 'contrastive' information in focus, whereas speakers take care not to highlight 'given' information, i.e., information which is present (explicitly or implicitly) in the preceding context. However, it is unclear whether such observations generalize to other languages as well; in addition, most studies on pitch accent distribution have not looked at the relation of such accents, which are encoded in the speech signal itself, to visual cues speakers may send to a communication partner. Therefore, in order to gain insight into the relative cue value of pitch accents for signaling focus, we will tackle this question in different perception tests, and approach it from (1) a *multilingual* and (2) a *multimodal* perspective. In the studies presented below we shall ignore newness accents, for reasons that will become clear later, and hence in the present study being contrastive is equivalent to being in focus.

First, we will present a cross-linguistic approach, given that the communicative importance of pitch accents is likely to be different for different languages. Consider the following utterances (after Cruttenden 1993) i.e., readings of the scores in English and Italian football reports (capitalized words indicate accents). In both examples, the football score is a tie so that a particular number (the words 'one' and 'uno') is repeated:

English
TOTTENHAM ONE - LIVERPOOL one

Italian
UDINESE UNO - ROMA UNO

In a typical English realization of such scores, the second instance of 'one' is deaccented, since it has just been mentioned in the preceding phrase. In the Italian scores, the second instance of 'uno' is typically accented again, even though it is literally given from the preceding context. This second accent in the Italian case is due to the fact that Italian strongly disfavours deaccentuation *within* NPs or other syntactic constituents (Ladd 1996:177-178 & p.c.). More general, Italian, as other Romance languages such as Catalan or French, is an example of what Vallduví (1992) would call a *non-plastic* language, i.e., a language which has a rather constrained intonation structure to mark information status, but which more heavily uses word order variation for that purpose. These languages are different from the *plastic* ones, such as Dutch and English, whose prosodic pattern is "moulded" to fit the information structure, so that intonation is used to mark information status. These claims do not entail that deaccentuation is impossible in Italian, as Ladd acknowledges that deaccentuation on sentence level in Italian is entirely possible, e.g., repeated full NPs may be deaccented (see also Avesani et al. 1995; Hirschberg & Avesani 1997; D'Imperio 1997), but they do mean that under certain conditions deaccentuation is infelicitous. The current study aims to test whether differences between accent structures in Italian and Dutch, as two cases of a non-plastic and a plastic language, respectively, have implications on the way listeners perceive focus.

Second, apart from variation between languages, the importance of pitch accents may also depend on the communicative setting in which they are used, in particular if we compare communicative settings in which dialogue participants can or cannot see each other during a spoken interaction. Different studies have suggested that there exist specific visual cues to focus structure as well. In particular, like pitch accents, rapid eyebrow movements have been claimed to play an accentuation role (e.g., Birdwhistell 1970, Condon 1976). It has even been argued that there is a one-to-one connection between the two; see, for instance, the so-called *Metaphor of Up and Down* (Morgan 1953, Bolinger 1985:202ff): when the pitch rises or falls, the eyebrows follow the same pattern. In fact, to see that there is indeed a close connection between pitch and eyebrows, one may try to utter a two word phrase, say "blue square", with a pitch accent (but no corresponding eyebrow movement) on the word "blue" and a rapid eyebrow movement (but no corresponding pitch accent) on the word "square". Most people find this a difficult exercise.

One of the few empirical studies devoted to the relation between pitch accents and eyebrow movements is Cavé et al. (1996), who report on a significant correlation between the two (in particular, and surprisingly, for the *left* eyebrow). It appears that rapid eyebrow movements often co-occur with pitch accents. The opposite is not the case: people do *more* with their pitch than with their eyebrows. Cavé and co-workers suggest that eyebrow movements and pitch do not link automatically (e.g., due to muscular synergy), but coincide for communicative

reasons. Naturally, this raises the question what these communicative reasons might be. In the literature on Talking Heads (i.e., combinations of computer animations with speech), there is no consensus on the timing and placement of eyebrow movements. Pelachaud et al. (1996) note that the decision to raise the eyebrows is affect dependent, but in the examples they discuss, pitch accents and eyebrows coincide. Thus to the question *I know that Harry prefers POTATO chips, but what does JULIA prefer?* the Talking Head of Pelachaud et al. (1996:19) would respond with the following utterance, in which capitalized words again indicate an accent, whereas overlined words are accompanied by a rapid eyebrow movement:

$$(\overline{\text{JULIA}}\ \text{prefers})_{\text{theme}}\ (\overline{\text{POPCORN}})_{\text{rheme}}$$

Cassell et al. (2001) use eyebrow raising (or "flashes" as they call them) more sparingly. The eyebrows are raised when an *object* in the "rheme" is described. So in reply to the question above, the algorithm of Cassell et al. would not produce a 'flash' on "Julia". It should be noted that neither Pelachaud et al. (1996) nor Cassell et al. (2001) report on evaluation: it is not known whether the animations are effective in the way human listeners process the information. We get no insight in the contribution of the eyebrow movement: its function remains unclear. Again, to learn more about the relative importance of pitch accents and eyebrow movements, this issue is tackled in the current study from a perceptual point of view, testing how listeners detect focus in audiovisual stimuli.

To facilitate comparisons across languages and across modalities regarding the cue value of pitch accents to signal focus, we have set up a particular experimental paradigm which can be applied to different languages and to both speech-only and audiovisual stimuli. The experiment consists of a perceptual task in which listeners essentially have to detect the main focus in an utterance. More specifically, subjects are instructed to decide, solely on the basis of a particular utterance, what the information would have been in the *preceding* utterance, i.e., subjects have to 'reconstruct the dialogue history'. Rather than using manipulated speech materials (read-aloud or synthetic) with controlled prosodic properties, the stimuli for the perceptual task discussed here consist of semi-spontaneous data whose intonational features are untouched when used in the test. By using naturally elicited speech materials, one avoids the risk that one tests the effect of intonation contours that are not representative of real data. For that purpose, we developed a specific dialogue game that triggers speakers' productions of different focus distributions in particular target sentences. The paradigm works for different languages so that it becomes easier to make cross-linguistic prosodic comparisons. In addition, the resulting utterances can be combined with visual cues, which makes it possible to study the relative cue value of accents and visual information.

In the next section, we first describe the experimental design to elicit accent patterns in both Dutch and Italian utterances, and the method to create audiovisual stimuli to be used in a series of perception tests. The following sections then describe the procedure and the results of the actual experiments on the perception of focus in speech-only stimuli in Dutch (study 1) and in Italian (study 2), and in

multimodal stimuli in Dutch (study 3). We end with a general discussion and a conclusion.

2. MATERIALS

2.1. Speech

For all three studies, utterances were used which were obtained in a semi-spontaneous way via a simple dialogue game. The game was played each time by two subjects, call them A and B, separated from each other by a screen. Figures 1 and 2 visualize the experimental set-up with a bird's-eye perspective on the starting situation of the game and the situation after the first turn in the game. In each game, both players have an identical set of eight cards at their disposal, each card showing a geometrical figure in a particular colour. Four of these cards are put on a stack in front of them, the four other cards are in a row before them. The four cards in the stack of A are the same as the four cards in the row of B, and vice versa. The game consists of a series of turns in which one participant gives instructions to select a card with a particular geometrical figure and the other follows these instructions. In each consecutive turn, the participants switch roles so that the original instruction-giver becomes the instruction-follower, and the other way around. In turn 1, the instruction giver, say A, begins with describing the figure on the top of his stack ("a blue square"). After he has described this figure, he removes it from his stack and puts it behind number 1 on his list. The instruction follower, B, listens to the description of A and removes that figure from his row of figures, and also puts it behind number 1 on his list. Now, the participants switch roles, so that B describes the figure that is on top of his stack ("a black triangle"), and A follows the instructions of B which will prompt both A and B to place the card with this object on the second place in the row with figures, and so on. The game is over when both players have no cards left. Each pair of subjects played a sequence of eight games, each time separated by a break of at least two minutes. Note that the players are given the instruction to describe the figure on top of their stack in terms of its colour and figure property. Speakers generally found it a very easy game to play, and as a consequence there are no faulty descriptions in the respective data sets.

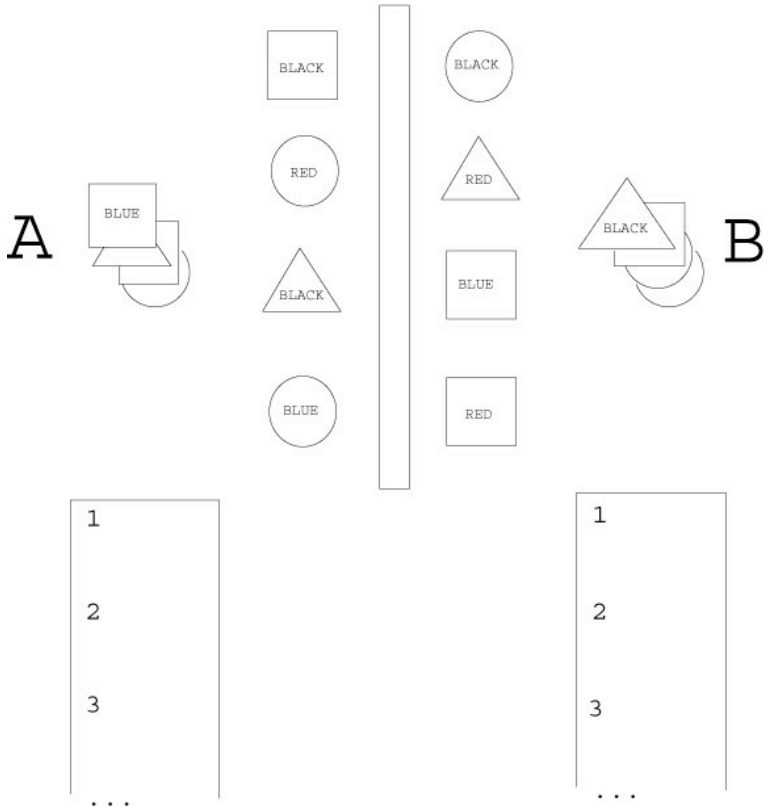

Figure 1: Visualization of the initial set-up of the experiment to elicit different referring expressions. A and B represent the two participants in the dialogue game. In the actual experiment, the different figures were given different colours. Further explanations in the text.

The speech data thus obtained allow for an unambiguous operationalization of the relevant contexts. A property is defined to be *new* (N) to the conversation if it is mentioned in the first turn of the current dialogue game, it is *given* (G) if it was mentioned in the previous turn and finally a property is *contrastive* (C) if the object described in the previous turn had a different value for the relevant property. We define a property to be in focus, if it is not given. (In the three studies described below, we will ignore newness and hence in these studies a property will be in focus if, and only if, it is contrastive.) By systematically varying the order of the cards in the stack, target descriptions (Dutch: "blauw vierkant" (blue square); Italian: "triangolo nero" (black triangle)) could be collected in all contexts of interest: no contrast (all new, NN), contrast in the prefinal word (CG), contrast in the final word (GC), all contrast (CC). Notice that in the 2-letter abbreviations, the first letter corresponds with the contextual status of the first word, and the second letter with the contextual status of the second word. Table 1 summarizes the situation. It is

worth noting that in the Dutch elicited utterances the adjective always precedes the noun, whereas in the Italian data it follows the noun. In other words, if we refer to the first word in the elicited NPs, we mean the adjective in case of the Dutch data, and we mean the noun in the case of the Italian data.

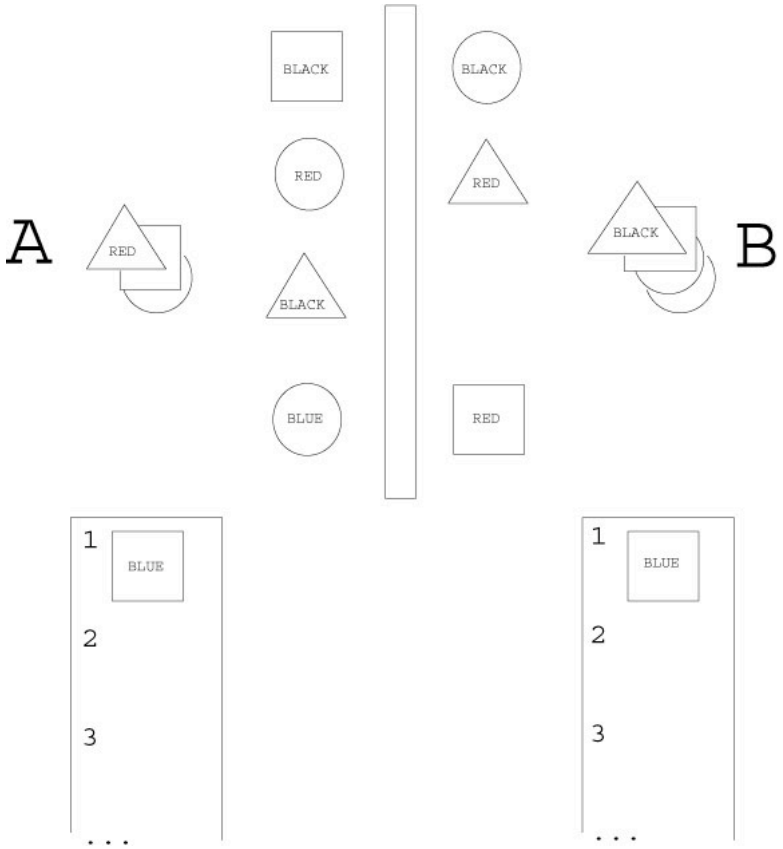

Figure 2: Visualization of the set-up of the experiment after A's first move ("blue square")

Eight Dutch speakers were recruited from students and colleagues from IPO, speaking the variant of standard Dutch as spoken in the Netherlands; the eight Italian speakers we recorded were all living in Italy, and were native speakers of the Tuscan variety of Italian. The Dutch speech materials are used in studies 1 and 3, the Italian ones in study 2.

Table 1: Examples of the four contexts.

NN		(beginning of game)
	B:	"blue square"
CC	A:	"red circle"
	B:	"blue square"
CG	A:	"yellow square"
	B:	"blue square"
GC	A:	"blue triangle"
	B:	"blue square"

2.2. Animations

For study 3 we combined the Dutch speech materials with an animated talking head. Since this was a male head, we only used the four male voices collected for Dutch. In addition, two synthetic male voices were used, copying the intonation contours of two of the human voices. We use both synthetic and natural voices in order to see to what extent the naturalness of the voice influences the perception of focus. A human voice has more natural and better sounding prosody, but a synthetic voice might be better suitable to accompany the visual counterpart of a synthetic character. A Dutch diphone speech synthesizer was used for the generation of the two synthetic versions. The animations were produced with the *CharToon* environment (Ruttkay et al. 1999). A 2D head of a male person formed the basis of the animations. Visual speech is generated on the basis of a set of 48 visemes (elementary mouth positions). Phonemes from the input are matched to corresponding visemes with a sampling rate of 100 ms, while intermediate stages are computed using linear interpolation. Rapid eyebrow movements coincide with the stressed syllable of either the first ("blauwe") or the second word ("vierkant"). Notice that these are the eyebrow counterparts of focus on the adjective and focus on the noun respectively. We did not include an eyebrow counterpart to "all focus", since this would involve either a raised eyebrow for a longer stretch of time or two rapid eyebrow movements in succession. For Dutch subjects both of these primarily have a non-focus signalling interpretation. It is worth stressing that in certain stimuli eyebrow movements are associated with words which are not accented. Eyebrow movements always had the following pattern: first, a 100 ms dynamic raising part, then a static raised part of 100 ms, and finally a 100 ms dynamic lowering part. The overall length of the movement is comparable to the average duration of rapid eyebrow movements of human speakers (± 375 ms, Cavé et al. 1996). We opted for slightly shorter movements due to the overall short duration of the stimuli. Figure 3 shows two stills from a typical animation used in the experiment.

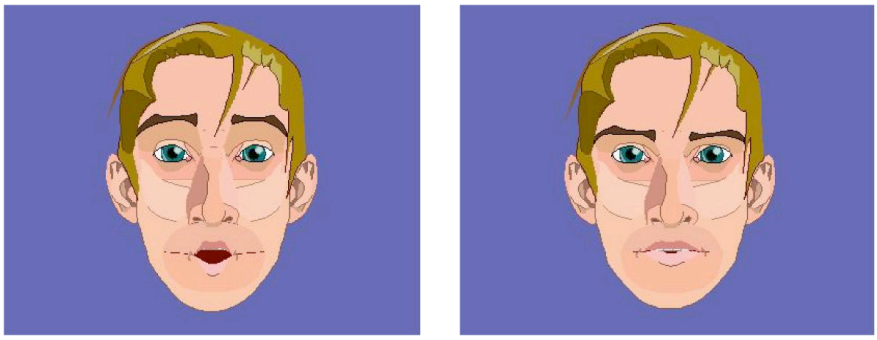

Figure 2: Two stills from the Talking Head uttering "blauw vierkant" (blue square) with a raised eyebrow on the first word (left) and no eyebrow action on the second word (right)

3. STUDY 1: FOCUS IN DUTCH

3.1. Preliminaries

The first study tests to what extent Dutch listeners are able to determine the main focus in an utterance by means of pitch accent distribution. For this purpose, we used data collected via the game described above. Before performing the dialogue reconstruction experiment, a distributive analysis of the target utterance "blauw vierkant" (blue square) was carried out. A consensus labelling was done by three independent intonation experts. The results of the labelling can be summarized as follows: in most cases, a property which is in focus receives a pitch accent. Interestingly the only exceptions to this general rule can be attributed to speaker differences among the eight speakers. One group of four speakers always end their utterance on a low boundary tone and always associate focused properties with a pitch accent. The four remaining speakers uniformly employ high boundary tones, and they associate the CC utterances with a single accent on the noun.

Table 2: Summary of the results of Study 1: classification of all 24 stimuli, for all 25 listeners (n=600). The vertical axis indicates the actual CONTEXT of the target utterance "blauw vierkant" (blue square). The horizontal axis indicates how many subjects CLASSIFIED the utterance in each of the three contexts.

		CLASSIFIED as			
		CC	GC	CG	Total
CONTEXT	CC	95	83	22	200
	GC	60	119	21	200
	CG	10	6	184	200

3.2. Procedure

Dialogue reconstruction data were obtained from 25 native speakers of Dutch (different from the eight speakers). The experiment was performed on an individual basis and was self-paced. All three versions (CG, GC and CC) of the target utterance ("blauw vierkant") produced by the eight speakers were used, making a total of 24 stimuli. In studies 1 and 3, Dutch subjects are presented with speech realizations of "blauw vierkant" taken from their original context, and the task is to determine by forced choice whether the preceding utterance would be: (1) "rood vierkant" (red square), (2) "blauwe driehoek" (blue triangle) or (3) "rode driehoek" (red triangle). The corresponding contexts are (1) CG (focus on the first word), (2) GC (focus on the second word) and (3) CC (all focus), respectively. The stimuli were presented in two random orders, to compensate for potential learning effects. Before the actual experiment started subjects entered a brief training session (consisting of three stimuli) to make them acquainted with the materials and the setting of the experiment. No feedback was given on the correctness of their answers, and there was no communication with the experimenters. Notice that the all new situation (NN) is not incorporated in the experiment, because there are no utterances preceding the NN so that subjects cannot reconstruct the preceding utterance. The NN utterances have been studied extensively in Krahmer & Swerts (2001), to find out whether there are prosodic differences between newness and contrastive accents in this setting[2].

3.3. Results

Table 2 contains the results for all eight speakers taken together. The overall distribution is significantly different from chance ($\chi^2 = 395.3$, df = 4, $p < 0.001$). The first thing to note is that for each line the highest numbers are on the diagonal. This means that each context is most likely to be classified correctly. However, these chances are much higher in the case of single focus, on contrastive items (CG and GC) than in the all focus case (CC). Subjects are particularly good in reconstructing the dialogue history when the adjective is the single focused item (note that these are the classic cases of narrow scope), which stands out prosodically due to the occurrence of a nuclear accent in non-default position. However, also when it is the noun that is the single item in focus, subjects are generally capable of reconstructing

[2] *Superficially, newness accents and contrastive accents appear to differ in our data, but a closer look reveals that this is not the case. In particular, at first sight it seems that (1) single contrastive items on the adjective (CG) have a different shape from newness accents in the same position and (2) contrastive items are judged to be more prominent than newness accents. However, (1) the difference in accent type is not so much associated with a contrast-specific prosodic shape but with the occurrence of a nuclear accent in a non-default position. And (2) the perceived prominence is not so much the result of inherent melodic properties of contrastive accents but seems due to the fact that the prosodic context does not contain other intonationally comparable pitch peaks. When the words are presented in isolation, contrastive accents are not perceived as more prominent than newness accents.*

the context. Interestingly, the number of confusions with the all focus (double contrast) context increases. This seems to imply that there is at least some amount of broad focus / narrow focus ambiguity (but see below), although the narrow focus interpretation is still prevalent. This result is compatible with earlier findings from Gussenhoven (1983) and Rump & Collier (1996) that these ambiguous cases are more confusable than the CG case, which only allows a narrow focus interpretation. In the case of double contrast there appears to be a very substantial broad vs. narrow focus confusion.

However, looking at the results for each speaker separately (all significantly different from chance as well), reveals an interesting difference between high and low boundary speakers. The main difference between speakers is found for the double contrast (CC) case. For low boundary speakers, utterances made in a CC context are predominantly classified as CC. Strikingly, this is not the case for high ending speakers, whose CC utterances are very frequently classified as GC utterances, which matches the earlier observation that these speakers tend to produce all-contrast utterances with a single accent on the noun. Thus, the fact that in table 1 CC utterances are often misclassified as GC utterances is essentially due to the difference between low and high ending speakers rather than broad vs. narrow focus interpretations.

4. STUDY 2: FOCUS IN ITALIAN

4.1 Preliminaries

The second study tests to what extent Italian listeners are capable to determine the main focus and reconstruct the dialogue history of an utterance using prosodic cues. Before performing the dialogue reconstruction experiment, a distributive analysis of the target utterances “triangolo nero” (black triangle) was performed. Three independent intonation experts listened to all realizations of “triangolo nero” produced by the eight speakers in the various contexts of interest, and decided on which words they perceived an accent. The three judges were in full agreement: every word is always accented, irrespective of context. All speakers produce the same contour, namely a flat hat shape with the second accent downstepped with respect to the first. Of course, it might be that different kinds of accents are realized in different contexts. However, an analysis of the fundamental frequency did not reveal any differences between contexts (see Swerts et al. 2002). In addition, we found no evidence for a clear correlation between information status and the perceived prominence of accents for the Italian data. Therefore, it seems a reasonable hypothesis that, contrary to the Dutch subjects, Italian subjects will not be able to reconstruct the dialogue history on the basis of prosodic cues.

4.2 Procedure

Subjects of the second dialogue reconstruction experiment were 25 native speakers of Italian (different from the eight speakers), mostly from Tuscany. The experiment was performed on an individual basis and was self-paced. All three versions (CG,

GC and CC) of the target utterance ("triangolo nero") produced by the eight speakers were used, making a total of 24 stimuli. In this study, Italian subjects hear versions of "triangolo nero" (black triangle), and have to guess whether the preceding utterance was (1) "rettangolo nero" (black rectangle), (2) "triangolo viola" (violet triangle) or (3) "rettangolo viola" (violet rectangle), again representing the following contexts: (1) CG (focus on the first word), (2) GC (focus on the second word) and (3) CC (all focus) respectively. The stimuli were again presented in two random orders, to compensate for potential learning effects. Before the actual experiment started subjects entered a brief training session (consisting of three stimuli) to make them acquainted with the materials and the setting of the experiment. No feedback was given on the correctness of their answers, and there was no communication with the experimenters.

4.3 Results

The results of the Italian reconstruction experiment on the basis of all eight speakers are displayed in table 3. A χ^2 analysis reveals that the distribution is *not* significantly different from chance. Looking at the results of the eight individual speakers, we see that the results for seven of them are not significant.[3] The picture is significantly different from the one obtained for the Dutch data (Pearson $\chi^2 = 223.8$, $df = 8$, $p < 0.001$). Thus, as expected, Italian listeners are not able to reconstruct the prior dialogue context on the basis of prosodic properties of the current utterance, in contrast to Dutch listeners.

Table 3: Summary of the results of Study 2: classification of all 24 stimuli, for all 25 listeners (n=600). The vertical axis indicates the actual CONTEXT of the target utterance "triangolo nero" (black triangle). The horizontal axis indicates how many subjects CLASSIFIED the utterance in each of the three contexts.

		CLASSIFIED as			
		CC	GC	CG	Total
CONTEXT	CC	52	70	78	200
	GC	53	82	65	200
	CG	61	73	66	200

5. STUDY 3: FOCUS IN AUDIO-VISUAL SPEECH

5.1 Preliminaries

In the third study we investigate the relative contributions of pitch accents and eyebrow movements to the perception of focus in Dutch. For this purpose, we use an

[3] *The results for the eighth speaker were just above the significance threshold. This was due to the fact that his CC utterance was often classified as CG. There is no obvious reason for this. Anyway, it is hard to see how this can be related to information status.*

animated male Talking Head and six different male voices. Four of these voices are human, and have also been used in study 1. The two remaining voices are synthetic, with the respective intonation contours copied from two of the human speakers. This makes it possible to compare the results of study 3 with those of study 1. The rapid eyebrow movements have been shown to be clearly perceivable. A further test indicated that the eyebrow movements boost the perceived prominence of words that also receive a pitch accent, and downscale the prominence of unaccented words in the direct context of the accented word (see Krahmer et al. 2002b). The question of interest to us here is whether this also has functional ramifications.

5.2 Procedure

A total of 25 native speakers of Dutch participated in the audio-visual dialogue history reconstruction experiment (different from the eight speakers, and also different from the 25 listeners from study 1). The experiment was individually performed and self-paced. Subjects watched and listened to the Talking Head uttering the two-word phrase "blauw vierkant" (blue square), with a particular intonation contour (taken from its original context; CG, GC or CC) and a rapid eyebrow movement on either the first or the second word. Eyebrow movements are indicated with a hat on the relevant item; the resulting six contexts are ĈG, CĜ, ĜC, GĈ, ĈC and CĈ. Since six voices are used the total number of stimuli is 36. The stimuli were displayed on a high-resolution color PC screen, sound came over the loudspeakers to the left and the right of the screen. Dutch subjects had to perform the same task as those of study 1, except that they were now presented with audiovisual stimuli. The stimuli were presented in two different random orders, to compensate for possible learning effect. Before the experiment started, subjects entered a brief training session (consisting of three stimuli) to make them acquainted with the material and the setting of the experiment. No feedback was given on the 'correctness' of their answers and there was no communication with the conductor of the experiment.

Table 4: Summary of the results of Study 3: classification of all 36 stimuli, for all 25 listeners (n= 900). The vertical axis indicates the actual CONTEXT of the target utterance "blauw vierkant" (blue square) plus the word which is associated with a rapid eyebrow movement. The horizontal axis indicates how many subjects CLASSIFIED the utterance in each of the three contexts.

		CLASSIFIED as			
		CC	GC	CG	Total
CONTEXT	ĈC	64	41	45	150
	CĈ	59	70	21	150
	ĜC	34	91	25	150
	GĈ	33	90	27	150
	ĈG	16	22	112	150
	CĜ	16	30	104	150

5.3 Results

Table 4 summarizes the results. The total distribution is significantly different from chance: $\chi^2 = 292.2$, df = 10, $p < 0.001$. First consider the cases with single pitch accents, i.e., the cases with a single prosodic focus on either the adjective or the noun. Notice that in these cases the majority of subjects indeed perceived the focus on the adjective or the noun respectively, no matter which of the words is accompanied by an eyebrow movement. Subjects are somewhat more likely to classify the cases with the prosodic focus on the adjective correctly than those with prosodic focus on the noun. Certainly for these single prosodic focus cases, the distribution of pitch accents is more important for the perception of focus than the placement of eyebrow movements. This is also reflected by the fact that in the post-experiment interview, all subjects indicated that they paid most (if not all) attention to information in the auditory channel. Nevertheless, there is an overall effect of eyebrow movements: the distribution obtained with an eyebrow movement on the first word is significantly different from the distribution with a movement on the second word ($\chi^2 = 19$, df = 8, $p < 0.025$). Closer inspection of table 4 reveals that this is primarily due to cases with a double pitch accent. If we compare the cases in which the first word (the adjective "blauw") is associated with a rapid eyebrow movement with the cases in which the first word is *not* associated with such a movement, we see that in the former case the focus is perceived on the first word in 45 instances, as opposed to 21 in the latter situation. And, conversely, when we compare the cases in which the second word (the noun "vierkant") is associated with a rapid eyebrow movement with the cases in which it is not, we see that in the former case 70 times a subject classified the noun as being in focus as opposed to only 41 times in the latter case. In other words, when the intonation contour provides less cues about the focus (since it contains two pitch accents), eyebrow movements have relatively more impact. Overall, the results for the four human voices are similar to the results for two synthetic voices, albeit that the effect of eyebrow movements is a bit (but not significantly) more pronounced for the synthetic ones. One subject explicitly indicated that she "trusted" the human voices more than the synthetic ones, and thus paid special attention to pitch accents in the former situation.

6. DISCUSSION AND FUTURE WORK

The perceptual approach to intonational phenomena has most strongly been promoted in the so-called IPO school of intonation ('t Hart et al. 1990). The original goal of this approach was to develop a *formal* metalanguage to describe the intonational properties of Dutch and a few other languages. Starting from the observation that perception acts as a filter that can stylize the acoustic signal, this enterprise has led to a phonetically explicit specification of a few basic intonational categories, i.e., a limited set of pitch rises and falls, that serve as building blocks out of which larger intonation contours can be constructed. In the current chapter, we have shown that such a perceptual approach is also useful to gain insight into *functional* aspects of intonation. In particular, we have shown that it helps to

comprehend how useful pitch accents are as signals of focus. This research question was tackled from a multilingual and multimodal perspective, applying a particular experimental approach, which consists of a dialogue game to elicit target utterances in different discourse contexts, and a series of perception tests to evaluate the functions of accents in different languages and in different communicative settings.

As to the results of the current study, we have found that the two languages investigated, Dutch and Italian, are markedly different regarding accent patterns inside NPs. In Dutch, it appears that accent patterns are indeed used to mark information status: accent distribution is the main discriminative factor with new and contrastive information generally accented, while given information is deaccented. Study 1 shows that our Dutch listeners are capable, in the majority of the cases, to reconstruct the prior dialogue utterance on the basis of properties of the current utterance. Italian differs from Dutch in terms of accent structure: distribution is not a significant factor in this language, since within the elicited NPs both adjective and noun are always accented, regardless of the information status. As a result, it is not surprising that the Italian listeners fail completely to interpret the target utterances in terms of the dialogue history (study 2). As noted in the introduction, Italian, being a *non-plastic* language, has other means besides prosody of marking information status. For instance, it has a freer word-order than *plastic* languages such as Dutch, and it is known to exploit this freedom to mark information status. However, the constraints of the experimental paradigm did not offer any room for Italian speakers to use word-order as an indicator of information status. Therefore it would be interesting to look for an experimental set-up in which speakers have more freedom to describe a particular state of affairs. This might also shed a different light on the deaccentuation debate, given that Ladd claims that deaccentuation of *complete* NPs within a sentence is quite possible in languages like Italian, which is supported by data from previous studies (Avesani, Hirschberg & Prieto 1995, D'Imperio 1997, Hirschberg & Avesani 1997).

Regarding the outcome of the audiovisual test (study 3), we have found that both auditory (accent distribution) and visual (eyebrow movement) cues can have a significant effect on the perception of focus. However, the effect clearly differs in magnitude; the impact of pitch accents is large, that of rapid eyebrow movements comparatively small. The visual cues contribute more when the auditory cues are inconclusive. Thus, for the condition which caused most confusion in study 1, eyebrows contribute the most in study 3. One consequence of the overall dominance of speech is that inconsistent cues go largely unnoticed (although a recent experiment indicates that subjects have a preference for animations in which eyebrow movements coincide with pitch accents, Krahmer et al. 2002b). That the auditory cues appear to be more important for focus perception may —with hindsight— be explained as follows: since human speakers do more with their pitch than with their eyebrows, it is not unnatural that human listeners have learned to pay more attention to changes in pitch than to eyebrow movements. It is interesting to compare the result of study 3 with those of study 1. Since the auditory cues dominate the visual ones, it is no surprise that the results basically confirm the speech-only results of study 1. Nevertheless, there is clearly more confusion in the audio-visual case. In part, the increase in confusion can be ascribed to the presence

of the eyebrow movements. Certainly, they account for much of the "confusion" in the cases with a double pitch accent. However, eyebrows cannot account for the slight increase in confusion for the cases with a single pitch accent. It might be that the mere addition of a visual channel leads to more confusion (compare Doherty-Sneddon et al. 2001).

As possible follow-up studies, it is useful to investigate real speaker behaviour in natural interactions to gain more insight into possible visual cues. For study 3, use was made of an analysis-by-synthesis technique, creating stimuli whose visual properties were systematically varied to learn more about the relative effect of this parameter on focus perception. While the manipulations were inspired by claims in the literature, it would be nice to supplement the current results with findings of observations on real speakers to see whether they indeed use eyebrow movements for signaling focus as suggested here, or whether these mainly signal other types of information, if any. It would also be highly interesting to see what happens with Talking Heads for non-Germanic languages such as Italian. As shown above, the results of study 2 reveal that Italian listeners systematically fail to correctly classify the Italian utterances in terms of dialogue history when confronted with speech-only stimuli. We are currently planning to do the dialogue reconstruction experiment with an Italian Talking Head lifting its eyebrows on either the first ("triangolo") or the second word ("nero"). We would expect that rapid eyebrow movements have more impact for the Italian head than for the Dutch one, since the auditory cues are less informative for Italian than for Dutch. This would be in line with one of the findings of study 3, that eyebrow movements become more important when pitch cues are less clear.[4]

REFERENCES

Avesani C. (1997), I toni della RAI. Un esercizio di lettura inton ativa, in: *Gli italiani trasmessi: la radio*, Firenze, Accademia della Crusca, pp. 659-727.

Avesani, C., Hirschberg, J., Prieto, P. (1995), The intonational disambiguation of potentially ambiguous utterances in English, Italian and Spanish, in: *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, pp.174-177.

Birdwhistell, R. (1970), *Kinesics and context*, University of Pennsylvania Press.

[4] *POSTSCRIPT (2004) Since the first version of this chapter was written (2002), both follow up studies mentioned in the discussion have been carried out. Swerts and Krahmer (2004) report on a production experiment in which subjects were asked to pronounce short utterances with one syllable marked for focus. When the audio-visual recordings were analysed, it was indeed found that subject may use eyebrow movements to signal focus, but various other cues were found of which head movement and visual articulatory emphasis were the strongest. Krahmer and Swerts (2004) describe a series of experiments with an Italian Talking Head. Contrary to our expectations, Italian subjects made* less *functional use of eyebrow movements than Dutch subjects. In general, we found a number of interesting differences between subjects' evaluation of Dutch and Italian Talking Heads, but all of these could be reduced to prosodic differences between the two languages.*

Bolinger, D. (1986), *Intonation and its Parts*, London, Edward Arnold.

ten Bosch, L. (1991), *On the structure of vowel systems. Aspects of an extended vowel model using effort and contrast*, PhD dissertation, University of Amsterdam.

Cassell, J., Vihjálmsson, H., Bickmore, T. (2001).BEAT: the Behavior Expression Animation Toolkit, *Proceedings of SIGGRAPH'01*, Los Angeles, CA, pp.477-486.

Cavé, C., Guaítella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R. (1996), About the relationship between eyebrow movements and F0 variations, *Proceedings of the Internation Conference on Spoken Language Processing (ICSLP)*, Philadelphia, pp.2175-2179.

Condon, W. (1976), An analysis of behavioral organization, *Sign Language Studies* **13**:285-318.

Cruttenden, A. (1993), The de-accenting and re-accenting of repeated lexical items, *Proceedings of the ESCA workshop on Prosody*, Lund, pp.16-19.

Doherty-Sneddon, G., Bonner, L., Bruce, V., (2001), Cognitive demands of face monitoring: Evidence for visuospatial overload, *Memory \& Cognition* **29** (7): 909-919.

Gussenhoven, C. (1983), Testing the reality of focus domains, *Language and Speech* **26**: 61-80.

't Hart, H., Collier, R., Cohen, A. (1990), *A Perceptial Study of Intonation: An experimental-phonetic approach to speech melody*, Cambridge: Cambridge University Press.

Hirschberg, J., Avesani, C. (1997), The role of prosody in disambiguating potentially ambiguous utterances in English and Italian, *Proceedings of the ESCA workshop on Intonation*, 1997, Athens, pp.189-192.

D'Imperio, M. (1997), Narrow Focus and Focal Accent in the Neapolitan Variety of Italian, *Proceedings of the ESCA workshop on Intonation*, 1997, Athens, pp.87-90.

Krahmer, E., Swerts, M. (2001), On the alleged existence of contrastive accents, *Speech Communication* **34**:391-405.

Krahmer, E., Swerts, M. (2004), More about brows, in: Zs. Ruttkay and C. Pelachaud (eds.), *Evaluating ECAs*, Dordrecht: Kluwer Academic Publishers.

Krahmer, E., Ruttkay, Zs., Swerts, M., Wesselink, W. (2002a), Pitch, Eyebrows, and the Perception of Focus, *Proceedings of Speech Prosody*, Aix-en-Provence, pp.443-446.

Krahmer, E., Ruttkay, Zs., Swerts, M., Wesselink, W. (2002b), Perceptual evaluation of audiovisual cues for prominence, *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, pp.1933-1936.

Ladd, D. (1996), *Intonational Phonology*, Cambridge: Cambridge University Press.

Morgan, B. (1953), Question melodies in American English, *American Speech* **2**:181-191.

Pelachaud, C., Badler, N., Steedman, M. (1996), Generating facial expressions for speech, *Cognitive Science* **20**:1-46.

Pierrehumbert, J., Hirschberg, J. (1990), The meaning of intonational contours in the interpretation of discourse, in: Cohen, P., Morgan, J., Pollack, M. (Eds.), *Intentions in Communication*, Cambridge MA: MIT Press, pp.342-365.

Pitrelli, J.F., Beckman, M., Hirschberg, J. (1994), Evaluation of prosodic transcription labeling reliability in the ToBI framework, *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Yokohama, Japan, pp.123-126.

Rump, H.H. and Collier, R. (1996), Focus conditions and the prominence of pitch-accented syllables, *Language and Speech* **39**, 1-15.

Ruttkay, Zs., ten Hagen, P., Noot, H. (1999), CharToon; A system to animate 2D cartoon faces, *Proceedings Eurographics*.

Steedman, M. (2000), Information Structure and the Syntax Phonology Interface, *Linguistic Inquiry*, **31** (4): 649-689.

Swerts, M., Krahmer, E. (2004), Congruent and incongruent audiovisual cues to prominence, *Proceedings of Speech Prosody*, Nara, Japan.

Swerts, M., Krahmer, E., Avesani, C. (2002), Prosodic marking of information status in Dutch and Italian: A comparative analysis, *Journal of Phonetics* **30** (4), 629-654.

Vallduví, E. (1990), *The Informational Component,* Ph.D. Dissertation, University of Pennsylvania.