Reconstructing dialogue history

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

Published in:
Proceedings of the 7th Eurospeech Conference on Speech Communication and Technology (Eurospeech), 2001, September 3-7, Aalborg, Denmark

Publication date:
2001

Reconstructing Dialogue History

Marc Swerts[*,†], *Emiel Krahmer*[†]

[*] CNTS, Center for Dutch Language and Speech, University of Antwerp, Belgium
[†] IPO, Center for User-System Interaction, Eindhoven University of Technology, The Netherlands
{M.G.J.Swerts/E.J.Krahmer}@tue.nl

Abstract

This paper deals with a perceptual analysis of accent structure in Dutch to see to what extent listeners are able to reconstruct information from the prior discourse context on the basis of prosodic properties of the current utterance. Using data collected in an earlier dialogue game experiment, subjects were asked to perform a perceptual task in which they had to try and reconstruct what the previous utterance was on the basis of input utterances with different accent patterns. Our results reveal that listeners are able to correctly guess the prior context for a significant number of cases, but that performance depends on the type of intonation contour of the input utterance.

1. Introduction

In Germanic languages such as Dutch and English, speakers exploit intonation to encode the status of the information they try to convey to their listeners. In particular, the distribution of pitch accents marks how utterances should be integrated in the larger discourse context: these accents serve to distinguish new or contrastive information from information which is given from the prior context. There has been some work into how pitch accents may have an effect on listeners' processing of incoming utterances. Nooteboom & Kruyt (1987) instructed subjects to rate the naturalness of utterance sequences with different accent patterns, and showed that listeners are sensitive to appropriate accent distributions. Terken & Nooteboom (1987) performed psycholinguistic experiments and found that people's reaction times are longer when given information is accented or when new information is deaccented. The current paper aims to gain more insight into how the listeners' interpretation of incoming speech is affected by the distribution of different types of accents. The work presented is also listener-oriented, but tackles some new problems compared to previous studies, and also differs from earlier work from a methodological point of view.

First, there is a debate in the literature about whether or not a distinction exists between so-called narrow focus accents, where the 'scope' of an accent is limited to the word on which it occurs, as opposed to broad focus cases, where the scope of the accent may include up to a whole sentence. Consider:

I bought a white POODLE

A narrow focus reading of this example leads to a contrastive interpretation of this utterance, meaning that the speaker bought a white poodle, and not any other white dog, like a cocker spaniel, which the addressee might have in mind. (Note that a single accent on "white" only gives a narrow focus, contrastrive reading.) In the broad focus case, the sentence could be a general answer to a question like "What did you do yesterday?", where all the information in the response is basically new.

It is unclear from the literature whether such ambiguities between narrow and broad focus cases are resolved by phonological properties of the accents. Some maintain that narrow focus, contrastive accents *are* formally different from other accents, either because the *type of accent* is different for the contrastive cases (e.g. Pierrehumbert & Hirschberg, 1990), or because they are more *prominent* (Bartels & Kingston, 1994). Others, however, maintain that contrastive accents do *not* exhibit specific intonation features (e.g. Bolinger 1986:342). To shed some light on this debate, we will tackle this question from a perceptual point of view, to see how listeners interpret utterances whose accent patterns —theoretically at least— allow both a narrow and a broad focus reading, compared to cases for which only a narrow focus interpretation is possible. In addition, little is known about the possible effect on perception of the larger contour in which an accent occurs. There are reasons to suspect that such a larger contour may have an impact on the way listeners interpret incoming utterances. For instance, Shimojima et al. (1999), in their study of repetitive utterances in Japanese, report that such utterances are differently processed in terms of information status depending on whether they end in a high or low boundary tone. Interestingly, in the data described by Krahmer & Swerts (2001) and which we used as a point of departure for the current study, we found a clear distinction between so-called high-ending and low-ending speakers. Hence we have an additional variable to explore in the current study, in terms of the effect the intonation contour has on interpretation.

Apart from the fact that the preception experiment described below addresses somewhat different questions, it is also new in that it uses a different methodology. The two studies introduced above basically start from constructed speech materials, either read aloud fragments by a reader experienced to produce specific intonation contours (Terken & Nooteboom, 1987) or sentences with synthetically generated intonation contours (Nooteboom & Kruyt, 1987). With both these types of data, there is a danger that one tests cases that are not representative for what happens in naturally occurring utterances. The aim of the current experiment, therefore, is to use naturally elicited speech data, whose accent patterns will not be manipulated. In the next section, we will first describe the data which will be used in the current study. Then, we will move to the description of the design and the result of a perception experiment, showing that listeners are indeed able to use prosody as a means to deduce the information status of words in terms of the dialogue history. We will end with a discussion and some concluding remarks,

2. Data

The data used in the dialogue reconstruction experiment described below were obtained in a production experiment described in detail in Krahmer and Swerts (2001). This production

Table 1: The 4 contexts for B's utterance "blue square".

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"

experiment consisted of a set of simple dialogue games played by four pairs of subjects, all native speakers of Dutch. During the game both participants had to describe differently coloured, geometrical figures (such as a blue square) on cards placed in a stack in front of them.

The data thus obtained allow an unambiguous operationalization of the relevant contexts. A property (colour or figure) 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. By systematically varying the sequential order of the cards in front of the subjects, target descriptions were collected for the eight speakers in four contexts: no contrast (all new, NN), contrast in the adjective (CG), contrast in the noun (GC), all contrast (CC). Table 1 summarizes the situation. Two questions were addressed in Krahmer and Swerts (2001): (1) which words receive an accent in which contexts and (2) are these accents formally different. Ad (1): All utterances of two target descriptions ("blue square" and "red square") were used for a distributional analysis performed by two intonation experts. Table 2 summarizes the results and reveals a clear trend: in the NN (no contrast/all new case) both adjective and noun are (nearly) always accented, and in most cases the same holds for the CC (double contrast) cases. When one item is given, while the other is contrasted (i.e., the CG and GC cases), the contrasted item generally is the only accented word. Even though both CG & GC, and NN & CC are strikingly similar, there are two exceptions. First, there is a complete lack of postnuclear accents in the CG case, while occasionally prenuclear accents on the adjective occur in the GC case. Second, CC differs from NN in that there are a number of utterances in the CC context with an accent only on the adjective or the noun. Looking at these exceptional cases revealed that in many cases the speaker made a contrast with his or her *own* last utterance (this was especially clear for the unambiguous narrow focus cases with the sole accent on the adjective), thereby ignoring their partners last contribution. Interestingly, all these "egocentric" speakers happen to end their utterances on a high (H%) boundary tone, whereas the other speakers uniformly employed low (L%) boundary tones.

Now to question (2), do the contrastive accents have a specific intonational shape? If one makes the common assumption that a single accent on the noun is ambiguous between a broad focus and a narrow focus reading, then one might expect that a contrastive accent manifests itself most clearly in the noun position. However, for none of our speakers does a comparison of a single contrastive accent on the noun (GC) with a newness accent on that same syntactic position reveal differences with respect to the type of accent. Interestingly, at first sight

Table 2: Accent distribution according to two intonation experts (exp_1 and exp_2) on all target utterances "blauw vierkant" and "rood vierkant" (*blue square* and *red square* respectively) in four contexts: NN (no contrast), CC (all contrast), CG (contrast only in adjective), GC (contrast only in noun). One CG (red square) utterance is missing.

	Accent on					
	Adj Only		Noun Only		Adj and Noun	
Context	exp_1	exp_2	exp_1	exp_2	exp_1	exp_2
NN	0	0	0	2	16	14
CC	3	3	2	4	11	9
CG	15	15	0	0	0	0
GC	1	1	11	11	4	4

the single contrastive accent on the adjective (CG) *is* of a different type than the newness accent on that same syntactic position. However, the single contrastive accent on the adjective is of the *same* type as the accent on the noun. Thus: the difference in type of accent is only apparent, since in the CG context the adjective is associated with a nuclear accent in a non-default position.[1] So, as far as *type* of accent is concerned there seems to be no difference between contrastive and newness accents. This does not necessarily mean that hearers are not able to distinguish the two. It might be, for instance, that contrastive accents stand out perceptually. To explore this, perceptual data were obtained from sixteen prosodically naive subjects who participated in an individually performed listening task. From the eight speakers in the production experiment two were selected: JR (a low-ending speaker) and WY (a high-ending speaker). The data were presented in two conditions: *complete* (entire utterances) and *isolated* (words). The *rationale* for these two conditions is the following: if a contrastive accent really stands out (i.e., is perceptually distinguishable from a more neutral, newness accent), then this should be a property of the accent itself, and thus this should hold *both* for the complete condition *and* for the isolated one. In the complete condition, subjects could hear the utterances as they were originally produced by speaker JR or WY. In the isolated condition, listeners were presented with one word, either the adjective or the noun, taken from the original utterance. The listener was instructed (for each condition) to select the member of the pair which he or she thought was the most prominent: in the complete condition, they were asked to focus on either the noun or the adjective and to determine by forced choice which of the pairs contained the most prominent one. In the isolated condition, they had to select (again by forced choice) the word which they judged to be the most prominent. No specific definition of prominence was given to the subjects.

The results can be summarized as follows: in the complete condition, single contrastive accents stand out as the most prominent ones, irrespective of the intonation contour (high- vs. low-ending) and irrespective of the place of the accent in the utterance (adjective vs. noun). Similarly, given items are always judged to be the least prominent, while the all new and double contrast cases lie in between the two extremes. It is striking that in the isolated condition a different picture emerges, in that, for instance, the single contrastive accents are no longer perceived as being the most prominent ones. This suggests that prosodic information from the entire contour plays a central role in the

[1] We refer to Krahmer and Swerts (2001) for a more detailed analysis and the associated sound files.

complete condition, whereas in the isolated condition hearers solely base prominence judgments on acoustic properties of the target word (in particular pitch and loudness).

3. Method

The perception experiment described above brought to light that listeners are sensitive to the accent structure of our elicited utterances since they perceive differences in prominence. This finding does not necessarily entail, however, that the accent structure of these utterances is also functionally relevant for listeners, for instance in that it steers the way they interpret the incoming utterances. In addition, only two speakers were used in the experiment (since performing all the pairwise comparisons with data from all eight speakers would lengthen the experiment too much), so that we do not have a guarantee that these first results generalize to data from all speakers. Therefore, we set up a listening experiment using data from all speakers with the explicit aim to test whether the accent structure of our target utterances had an impact on the way listeners process these utterances semantically. Given that we found that speakers of Dutch encode the prior discourse context in the accent structure of the current utterance, our general question was whether listeners are able to reconstruct the dialogue history from such prosodic cues. More specifically, they were instructed to determine solely on the basis of a particular input utterance what the information was in the utterance preceding the current one. If they were able to do so, we were interested to find out whether this interpretation is dependent on different accent distributions, in particular for cases that allow both a broad and narrow focus interpretation, and those for which only a narrow focus reading (contrastive) is possible. In addition, we wanted to explore whether or not the data from our two speaker types, the high-ending and low-ending ones, were differently processed.

Dialogue reconstruction data were obtained from 25 native speakers of Dutch, a minority of which is involved with speech research. The experiment was performed on an individual basis and was self-paced. The stimuli used in the experiment was one of the target utterances ("blauw vierkant", blue square) collected in the production experiment as described above. In the current experiment subjects were presented with these utterances of "blauw vierkant" taken from their original context, and the task was to determine by forced choice what the preceding utterance was: (1) red square, (2) blue triangle or (3) red triangle. The corresponding contexts are CG (contrast in the adjective), GC (contrast in the noun) and CC (all contrast respectively).[2] All versions of the target utterance "blauw vierkant" produced by the eight speakers in the production experiment were used, making a total of 24 stimuli. Before the actual experiment started, subjects entered a brief training session (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 conductors of the experiment. The entire experiment lasted approximately 5 minutes. Subjects could listen to each stimulus as often as they desired, although not much use was made of this option. The stimuli were presented in two different randomized lists, to compensate for potential learning effects.

[2] Notice that the all new situation (NN) was not incorporated in this experiment, because (1) the NN utterances are not contrastive since they are not uttered in the context of another description and (2) the accent distribution and the prominence experiment indicate that the NN and CC utterances are essentially indistinguishable from a prosodic perspective.

Table 3: Classification of utterances of all 8 speakers. The vertical axis indicates in which CONTEXT the target utterance "blauw vierkant" (*blue square*) was actually uttered, the horizontal axis indicates how many subjects CLASSIFIED the utterance in each of the three possible contexts. $N = 600$ (25 subjects $\times$ 8 speakers $\times$ 3 target utterances).

		CLASSIFIED AS			
		CC	GC	CG	*total*
CONTEXT	CC	**95**	83	22	200
	GC	60	**119**	21	200
	CG	10	6	**184**	200
	total	165	208	227	600

4. Results

All subjects made between 6 and 12 incorrect classifications, with 8 being the average number of errors. There was no effect of occupation; speech researchers did not score better than other subjects (in fact, the one person who made 12 errors is an internationally renowned speech expert). Table 3 contains the results for the data of all eight speakers. The resulting distribution is significantly different from chance (χ^2 = 395.3, df = 4, $p < .001$). One can read this table as a confusion matrix. The first thing to note is that for each row the diagonal contains the highest numbers. This means that each context is most likely to be classified correctly. However, these chances are much higher in the case of single contrastive contexts (CG and GC) than in the double contrast case. Subjects are particularly good in reconstructing the dialogue history when the adjective is the single contrastive item (the classic case of narrow scope), which stands out prosodically due to the occurrence of a nuclear accent in a non-nuclear position. However, also when the noun is the single contrastive item subjects generally are capable of reconstructing the context. Interestingly, the number of confusions with the CC context increases. This implies that there is at least some amount of broad focus / narrow focus ambiguity, although the narrow focus interpretation is still prevalent. In the case of double contrast there appears to be a very substantial broad vs. narrow focus confusion. However, closer inspection of the data reveals that here an interesting difference appears between high and low boundary speakers.

Tables 3 and 4 give the results for low and high boundary speakers respectively. The distribution for low-ending speakers is significantly different from chance (χ^2 = 232.9, df = 4, $p < .001$), as is the distribution for the high-ending speakers (χ^2 = 251.3, df = 4, $p < .001$). More interestingly, the respective distributions for low- and high-ending speakers are significantly different from each other (Pearson χ^2 = 73.8, df = 8, $p < .001$). The classifications for the single contrastive cases are more (CG) or less (GC) identical. The differences between high- and low-ending speakers in the GC case are largely due to one low-ending speaker whose GC utterance is often misclassified as CG. The main difference between the two kinds of 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. Thus, the fact that in table 3 CC utterances are often misclassified as GC utterances is essentially due to the difference between low- and high-ending speakers.

Table 4: Classification of utterances of the 4 speakers who end their utterance with a low boundary tone (L%). The vertical axis indicates in which CONTEXT the target utterance "blauw vierkant" (*blue square*) was actually uttered, the horizontal axis indicates how many subjects CLASSIFIED the utterance in each of the three possible contexts. $N = 300$ (25 subjects $\times$ 4 speakers $\times$ 3 target utterances).

		CLASSIFIED AS			
		CC	GC	CG	*total*
CONTEXT	CC	**62**	18	20	100
	GC	31	**49**	20	100
	CG	5	1	**94**	100
	total	98	68	134	300

Table 5: Classification of utterances of the 4 speakers who end their utterance with a high boundary tone (H%). The vertical axis indicates in which CONTEXT the target utterance "blauw vierkant" (*blue square*) was actually uttered, the horizontal axis indicates how many subjects CLASSIFIED the utterance in each of the three possible contexts. $N = 300$ (25 subjects $\times$ 4 speakers $\times$ 3 target utterances).

		CLASSIFIED AS			
		CC	GC	CG	*total*
CONTEXT	CC	33	**65**	2	100
	GC	29	**70**	1	100
	CG	5	5	**90**	100
	total	67	140	93	300

5. Discussion and conclusion

In general, listeners in the current experiment are able to reconstruct the dialogue history for a sizeable amount of the utterance stimuli, solely on the basis of the intonation pattern of the utterance. This effect is particularly strong in the case of single contrasts. In the double contrast case, the number of confusions is the largest. Closer inspection of the data reveals that this is largely due to differences between high- and low-ending speakers. More in particular, the CC utterances of high-ending speakers are likely to be classified as GC utterances. In fact, this is readily explained by the overall intonation patterns. It appears that our high-ending speakers tend to provide CC cases with a single accent on the noun, whereas our low-ending speakers use double accents in such cases (cf. Table 2 plus the description of this table in section 2). This difference in tendency to have the first word accented or not may be due to the type of the final boundary tone, and could be explained by the fact that speakers try to maximize the pitch difference between the onset and the closure of an intonation contour. It might be argued that, in case of low-boundary tones, speakers tend to produce an initial accent to have a high onset which makes a clear melodic contrast with the final boundary tone, whereas the absence of an accent results in a low onset that gives a clear melodic difference with the final high. Anyway, these speaker differences in terms of accent distribution match the bias in our perceptual results, with a preference for CG for low-ending speakers (134 out of 300 classifications), and a preference for GC for high-ending speakers (140 out of 300 classifications). Also, our data suggest some effect of speaker convergence, and a corresponding effect on listener results. Our original production data revealed that our speakers tend to copy the intonational strategies from their respective speaking partners, in that we only find high-ending speakers to communicate with other high-ending speakers, the same being true for the low-ending speakers. Accordingly, the perceptual results are similar for data from speakers who were conversation partners in our original dialogue game: good or bad classification results for one speaker tend to match the good or bad results for data from his/her partner. Whether this result is a coincidence or just an experimental artefact, or whether it is a true result that generalizes to naturally occurring conversations, is an interesting research question to be explored in the future. Note also that these data hold for Dutch, which has a relatively flexible prosodic structure. We have started to run exactly the same perception experiments with Italian data, using utterances collected with the same dialogue paradigm (Swerts et al., 1999). Italian has been claimed (e.g., Ladd 1996) to be intonationally different from languages such as Dutch and English, in that it strongly resists deaccentuation within particular syntactic constituents, such as in the NPs we elicited. Our acoustic data indeed reveal this to be the case, showing that our speakers always put an accent on both the adjective and the noun, irrespective of prior context, although we did find some phonetic differences in type of accent. Not unexpectedly, our first perceptual results from a pilot test reveal that listeners mostly fail to correctly classify these utterances in terms of dialogue history, which confirms the claim that Italian and Dutch are intonationally different in terms of accent structure.

6. References

[1] Bartels, C. and J. Kingston, "Salient pitch cues in the perception of contrastive focus", In: P. Bosch and R. van der Sandt (eds.) Focus and natural language processing, IBM working papers pp. 11–28, 1994.

[2] Bolinger, D., Intonation and its parts, Palo Alto, CA: Stanford University Press, 1986.

[3] Krahmer, E. and M. Swerts, "On the alleged existence of contrastive accents", Speech Communication, 2001 (in press).

[4] Ladd, D.R., Intonational phonology. Cambridge: Cambridge University Press., 1996.

[5] Nooteboom, S.G. and J.G. Kruyt, "Accents, focus distribution, and the perceived distribution of given and new information: an experiment", JASA 82, 1512-1524, 1987.

[6] Pierrehumbert, J. and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse". In: P.R. Cohen, J. Morgan and M.E. Pollack (eds.), Intentions in communication, Cambridge: MIT, pp. 271–311, 1990.

[7] Shimojima, A., Y. Katagiri, H. Koiso and M. Swerts, "An experimental study on the informational and grounding functions of prosodic features of Japanese echoic responses", Proc. ESCA Workshop on Dialogue and Prosody, Veldhoven, 187-192, 1999.

[8] Swerts, M., C. Avesani and E. Krahmer, "Reaccentuation or deaccentuation: a comparative study of Italian and Dutch", Proc. 14th ICPhS, San Francisco, 1541-144, 1999.

[9] Terken, J. and S.G. Nootenboom, "Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information", Language and Cognitive Processes 2 (3/4), 145-163, 1987.