Now move X to cell Y: Intonation of 'now' in on-line reference resolution

Bettina Braun and Aoju Chen

Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
{bettina.braun;aoju.chen}@mpi.nl

Abstract

Prior work has shown that listeners efficiently exploit prosodic information both in the discourse referent and in the preceding modifier to identify the referent. This study investigated whether listeners make use of prosodic information prior to the ENTIRE referential expression, i.e. the intonational realization of the adverb 'now', to identify the upcoming referent. The adverb 'now' can be used to draw attention to contrasting information in the sentence. (e.g., 'put the book on the bookshelf. Now put the pen on the bookshelf.'). It has been shown for Dutch that *nu* ('now') is realized prosodically differently in different information structural contexts though certain realizations occur across information structural contexts.

In an eye-tracking experiment we tested two hypotheses regarding the role of the intonation of *nu* in online reference resolution in Dutch: the "irrelevant intonation" hypothesis, whereby listeners make no use of the intonation of *nu*, vs. the "linguistic intonation" hypothesis, whereby listeners are sensitive to the conditional probabilities between different intonational realizations of *nu* and the referent. Our findings show that listeners employ the intonation of *nu* to identify the upcoming referent. They are mislead by an accented *nu* but correctly interpret an unaccented *nu* as referring to a new, unmentioned entity.

1. Introduction

Prosodic information (i.e. pitch variation and phrasing) helps the listener to process utterances more efficiently than when this information is absent. For instance, the written words 'Peter was in London' do not tell the reader whether the writer intended to contrast 'Peter' with someone else. To do so unambiguously, a different syntactic construction has to be used (e.g., 'It was Peter who was in London'). In spoken language, the intended contrast is signaled by intonation, irrespective of word order (at least in English and Dutch).

Recent research has shown that listeners rapidly exploit such prosodic information in speech comprehension. Dahan and colleagues [6] conducted an eye-tracking experiment in which listeners were asked to follow two consecutive instructions to move an object on the computer screen. The screen depicted line drawings of two referents with an overlapping first syllable (e.g., 'candle' and 'candy'), and two unrelated distractors. The target words were either accented (signaling new or contrastive information) or not (signaling old or given information). Results revealed an initial bias towards the referent that was not mentioned in the first instruction. But upon hearing the first syllable of the target word, participants fixated the contrastive referent more when it was accented than when it was deaccented. That is, hearing an accented "CAN" in the second instruction leads to more fixations to the object not mentioned in the first instruction. Thus by exploiting prosodic information in the segmentally ambiguous first syllable of the target word (e.g., 'can'), listeners got a head start in resolving the ambiguity. That is, they were faster in identifying the intended referent than would have been possible if they had relied on segmental information alone.

Likewise, the prosodic realization of adjectives in the referring expression is exploited by listeners. Using the same method, Weber et al. [12] recorded instructions that contained an accented or unaccented color adjective in the second instruction (e.g. 'Click on the purple scissors. Now click on the red scissors/vase'). Results showed that participants tended to interpret the adjectival modification in the second instruction contrastively (i.e. more looks to red scissors than to red vase), but this tendency was enhanced when the adjective was accented. This suggests that prosodic information prior to the referent itself (though within the referential expression) helps the listener to identify the upcoming referent.

In the present study, we tested whether listeners also make use of prosodic information prior to the ENTIRE referential expression to identify the upcoming referent. A test case for this is the temporal adverb 'now', which is generally used in the above-mentioned experimental designs with two consecutive instructions and also occurs frequently in natural interactions [8]. In consecutive instructions, it signals a contrast to the first instruction. For instance, an instruction such as 'Move the candle above the triangle' can be followed by 'Now move the candy above the triangle' (contrast in referent) but also by 'Now move the candle above the square' (contrast in location), and also by 'Now move the candy above the square' (double contrast). In some languages, such as Dutch and German, the ambiguity in the locus of the upcoming contrast can be resolved to some extent by placing the adverb before the contrasting constituent (lit. 'Move the candle now above the square'). But this is not mandatory. In a prior production experiment in Dutch, in which speakers described video clips depicting the three kinds of contrasts, scrambling of the adverb was observed in 10% of the cases only [4]. Instead, to disambiguate the locus of the upcoming contrast, speakers varied the accentual realization of the adverb: *nu* ('now') was accented frequently (typically with H* and H*L) when the location was contrasted (96%) or when there was a double contrast (88%) but considerably less frequently accented when the referent was contrasted (62%). From these results we can calculate the probability of a contrastive or previously mentioned referent given an accented or unaccented *nu* using Bayes' theorem:

$$p(contrREF \mid accNU) = \frac{p(accNU \mid contrREF) * p(contrREF)}{p(accNU)}$$

Applying this formula, the probability of encountering a contrastive referent (contrast in location or double contrast) is 60% upon hearing an accented *nu* and 93% upon hearing an unaccented *one*. These conditional probabilities indicate a bias

towards a contrastive referent independent of the intonation of *nu*. It is thus possible that listeners make no use of the intonational information on *nu* in predicting the upcoming referent. We refer to this as the "irrelevant intonation" hypothesis. This would be in line with recent findings by Ito & Speer [9], who reported that the intonation of discourse markers such as 'and then', 'and next', 'after that' were not predictive of an upcoming contrast in either the adjective or the noun of adjective-noun pairs ('and next hang the green ball').

On the other hand, the probability of a contrastive referent is much higher following an unaccented *nu* (93%) compared to an accented *nu* (60%). Further, the probability of encountering the same referent (contrast in location) is substantially higher upon hearing an accented *nu* (39%) than upon hearing an unaccented *nu* (7%). Consequently, it is still likely that listeners make use of the intonational realization of the adverb to predict the upcoming referent. We refer to this as the "linguistic intonation" hypothesis.

We tested these two conflicting hypotheses regarding the interpretation of *nu*. The "irrelevant intonation" hypothesis predicts that listeners interpret the presence of the adverb as a cue for a contrastive referent, irrespective of its intonational realization. The "linguistic intonation hypothesis" predicts that listeners are sensitive to the conditional probabilities between the adverb and the referent such that they associated an unaccented adverb with a contrastive referent but an accented adverb with the referent mentioned before.

2. Methods

We conducted an eye-tracking experiment similar to the ones reported in Dahan et al. [6] but with three adjustments. First, to increase the number of word pairs that we could use as experiment materials, we replaced the black-white line drawings with printed words (see [11] for a discussion of this method). Second, as every object other than the referent mentioned in the first instruction was potentially contrastive, to sharpen the contrast we changed the standard display with four objects and four geometric shapes into a display with two objects (i.e. the referent mentioned in the first instruction and the contrasting one) and two geometric shapes (see Figure 1). Third, to increase the window in which anticipatory eye movements can be observed, we delayed the disambiguating information from the target by inserting the padding *het woord* 'the word' before the referent.

Participants' eye movements were recorded while participants followed pre-recorded instructions to move two objects.

2.1. Materials

Twenty-four disyllabic Dutch word pairs with stress on the first syllable were selected. All of them had an identical initial consonant-vowel sequence (e.g., *zegel-zetel, panda-panter*). One word in each pair was assigned the role of first referent, the other word the role of contrastive referent. Since a more frequent word (e.g., 'bed') attracts more looks than a phonologically overlapping but less frequent word (e.g., 'bell' [5], the words in each pair were matched for lexical frequencies according to the CELEX word form dictionary [1] (9.0 per million vs. 9.1 per million).

The first referent and the contrastive referent, together with a square and a triangle were displayed on a 5 × 5 grid on a computer screen (Figure 1). The words were shown in boldface black Arial 24 against a white background (96 × 96 pixels).

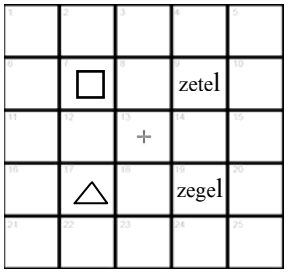

Figure 1: *Example display of a trial.*

Each trial consisted of two consecutive instructions to move an object in the display. The first instruction referred to the first referent (e.g. *Verplaats het woord zegel naar vak 21* 'Move the word stamp to cell 21'); the second instruction referred either to the first referent again (LOCATIONCONTRAST, e.g. *Verplaats nu het woord zegel naar vak 11* 'Now move the word stamp to cell 11') or to the contrastive referent (OBJECTCONTRAST, e.g. *Verplaats nu het woord zetel naar vak 21* 'Now move the word seat to cell 21').

The instructions were recorded by a female native speaker of Dutch who had attended intonation classes. She was given specific descriptions as to what intonation pattern to use (referring to ToDI, cf. Gussenhoven [7]). The first instruction was spoken with a falling initial boundary (%HL), H*L pitch accents on the referent and the location, and a low boundary tone. In the LOCATIONCONTRAST condition, the second instruction was produced with a prenuclear H*L accent on the adverb *nu* (which would be represented as L*H in ToBI), a deaccented target word and a nuclear H*L pitch accent on the location (Figure 2, upper panel). In the OBJECTCONTRAST condition, the adverb *nu* was deaccented, followed by a target with an H*L accent and a deaccented location (Figure 2, lower panel). The second instructions always started and ended with a low boundary tone.

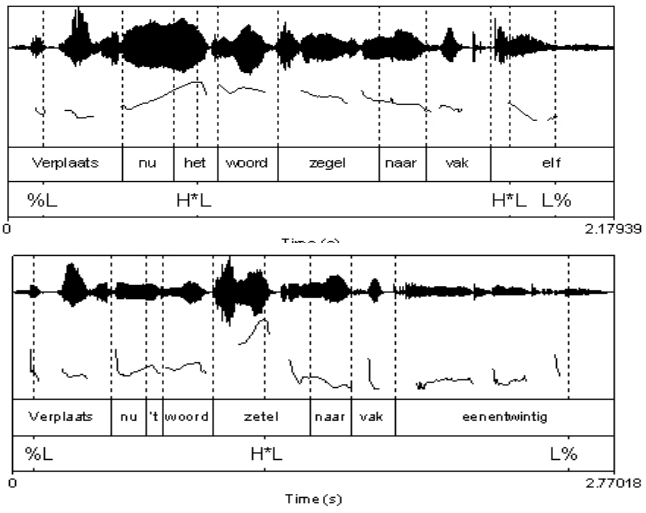

Figure 2: *Intonation of an instruction in LOCATIONCONTRAST with an accented nu (upper panel) and OBJECTCONTRAST with an unaccented nu (lower panel).*

Additionally, 34 filler trials were used, 18 of which were similar to the experimental trials but the two depicted words were phonetically unrelated (9 with a LOCATIONCONTRAST, and 9 with an OBJECTCONTRAST). To keep participants attentive, there was an additional set of 16 filler trials with only a single instruction, half of which contained word pairs similar to the experimental items. Participants were asked to click on a word or to move a word above or below the rectangle or the triangle.

Two master lists and two complementary lists were constructed, each containing 24 experimental trials. The word pairs were split into two groups, maintaining a matched mean frequency for first and contrastive referents. In master list A, the first half of the word pairs were assigned to the OBJECTCONTRAST condition, the other half to the LOCATION-CONTRAST condition. The same was done for master list B. But here the order of first (e.g., *zegel*) and contrastive referent (e.g., *zetel*) was swapped to minimize a potential bias for one of the two words in each pair; the locations of the words on the screen were also swapped (e.g. *zegel* in cell 9, *zetel* in cell 19 vs. *zegel* in cell 19, *zetel* in cell 9 in master list A). In the two complementary lists, every pair that was assigned to the OBJECTCONTRAST condition in the master lists was assigned to the LOCATIONCONTRAST condition, and vice versa. Twelve of the 34 filler items were placed at the start of all lists. The remaining filler trials were interspaced with the experimental trials. There were three randomizations for each of the four lists, resulting in 12 experimental lists.

2.2. Participants

Twenty-four native speakers of Dutch took part in the experiment for a small fee. They were naïve with respect to the purpose of the experiment. The experiment lasted about 30 minutes.

2.3. Procedure

Participants were randomly assigned to the experimental lists and were tested individually. They were first given written instructions on the task, and were then seated in front of a computer screen at a comfortable distance. An SMI Eyelink II eye-tracking system was fitted and calibrated. At the start of each trial, the two words and the two geometric shapes were displayed in cells 7, 9, 17, and 19 of the grid. Their positions were counterbalanced across conditions, so that each of the words and shapes occurred equally often in each position for each condition. Auditory stimuli were presented binaurally over headphones. The first instructions started simultaneously with the display of the grid. The second instructions started after the word mentioned in the first instruction was dropped into its new cell (but not before the end of the first instruction). An automatic drift correction was initiated after each block of six trials.

Participants' eye movements and mouse actions were monitored during the second instruction. The center of the pupil was tracked to determine the position of the eye relative to the head. Onset and offset as well as the coordinates of the fixations were recorded with a sampling rate of 250Hz.

3. Results

Only fixations before the mouse click to drag and drop the word were included in the analysis. Euclidean distances were calculated between a fixation and the center of each cell with an object (a referent or a geometric shape). If the fixation fell within the smallest circle surrounding the cell with an object (distance 67.6 pixels from the middle of the cell), it was counted as a fixation to that object. Fixations in the cross-section of two circles were assigned half to each of the corresponding cells.

Fixation proportions to each referent were calculated by dividing the number of fixations to that referent by the total number of fixations in 10ms steps. A visual summary of the evolution of fixation proportions to the contrastive referent over time is presented in Figure 3. Fixation proportions to the first referent constitute a near mirror image and are therefore not included in the display. Fixations to the triangle and square account for less than 3% of the total number of fixations and do not change over time; they are not included in the figure. The black line indicates fixation proportions in conditions with an accented *nu* (LOCATIONCONTRAST) – hereafter accented-nu condition, while the gray line shows fixation proportions in conditions with an unaccented *nu* (OBJECTCONTRAST) – hereafter unaccented-nu condition. The solid vertical lines indicate the averaged acoustic end point of the adverb *nu* and the start of the referent. Both landmarks are later in the accented-nu condition.

It takes typically about 150 to 200 ms to launch a programmed eye movement (e.g., [10]). Previous eye-tracking studies with visual objects have referred to the upper bound. Since we used two printed words instead of four objects, we adopted the more conservative estimate of 150 ms. Therefore, fixations caused by certain acoustic information can only be observed 150 ms after it.

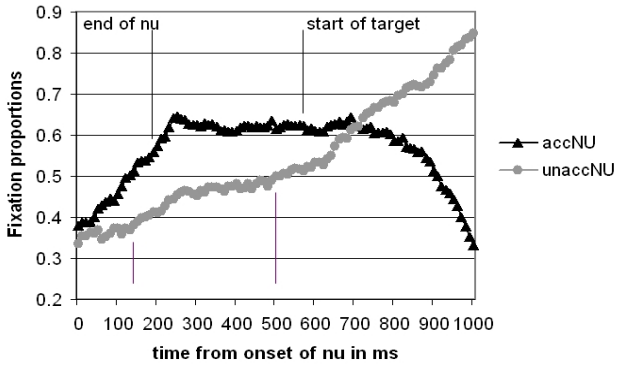

Figure 3: *Averaged fixation proportions to the contrastive referents in conditions with an accented and unaccented nu.*

For statistical analysis, fixations were recoded in 10ms steps as either pertaining to the cell of the contrastive referent or not; this allowed us to represent the categorical nature of the dependent variable (cf. Barr [2]). We then calculated multi-level logistic regressions with time and accentuation as fixed factors and participants and items as crossed random factors for different time windows (cf. Bates & Sarkar [3]). The size and direction of effects for significant predictors are indicated by the regression coefficients (βs). The exponent of the β value (e^{β}) serves as a measure of the increase in fixations to the contrastive referent.

Prior to the point that fixations could be driven by the acoustic input from the adverb (0-150 ms after onset of *nu*), there was a significant effect of accentuation (β=0.68, z=6.7, p<0.001), a significant effect of time (β=0.044, z=5.6, p<0.001), and an interaction between condition and time (β=0.023, z=2.18,

p<0.001): Fixations increased more quickly for the accented-nu condition than for the unaccented-*nu* conditions.

We then analyzed fixations in the ambiguous time window, i.e. when information about the realization of the adverb was processed but prosodic information about the referent was not yet available (from 350-720ms for the accented-nu condition and from 280-650ms for the unaccented-nu condition). As can be seen in Figure 3, the initial increase in the proportions of fixations to the contrastive referent continued steadily in the unaccented-*nu* condition (gray line). This is in line with the prediction from the "linguistic intonation" hypothesis that listeners interpret an unaccented *nu* as referring to a new referent. On the other hand, the initial increase in the proportions of fixations to the contrastive referent was 'frozen' in the accented-*nu* condition, i.e. changing little over time until the prosodic information about the referent (accented or unaccented) became available (at 720ms). These results appear to suggest that listeners interpret an accented *nu* as pointing to a contrast in referent too, contra our prediction derived from the "linguistic intonation" hypothesis that listeners interpret an accented *nu* as referring to the referent mentioned before. This may be explained by the fact that speakers also frequently accent *nu* when both the referent and the location are contrasted. Even though we only used the LOCATIONCONTRAST CONDITION and the OBJECTCONTRAST condition in our experiment, listeners might entertain the possibility of a double contrast. In this sense, fixating the contrastive referent is a useful heuristic for the task.

Statistical analyses confirmed these observations. For these analyses – due to variability in durations across conditions – the fixations in the ambiguous time window were resampled into 37 time units of approximately 10ms (if the actual duration of the ambiguous time window was longer than the average 370ms, the 37 time units contained information from slightly more than 10ms). We found a significant effect of accentuation (β=1.68, z=30.2, p<0.001). Furthermore, there was an interaction between condition and time (β=0.013, z=5.3, p<0.001), indicating that fixations increased more rapidly for the accented-nu condition than for the unaccented-nu conditions. Separate analyses on fixations to the contrastive referent in the two conditions showed that there was no effect of time in the accented-nu condition (z<1), but in the unaccented-nu condition fixation proportions increased over time (β=0.015, z=8.7, p<0.001).

Taken together, these results show that listeners are sensitive to intonational information prior to the entire referential expression. Our findings stand in contrast to a recent experiment by Ito and Speer [9], who found no anticipatory effects of intonation for the discourse markers 'and then', 'and next', 'after that' in signaling a contrast in either the adjective or noun of an adjective-noun pair. It is conceivable that the contrast locations in their study were too close (i.e., within a constituent rather than on different constituents). Alternatively, 'now' might be more effective in signaling changes in the information flow than these discourse markers.

4. Conclusion

We investigated the effect of intonation on reference resolution in Dutch. In contrast to earlier studies, we manipulated the intonation of a word outside the referential expression, i.e. the adverb *nu* ('now') in sentences such as "Now put the book on the bookshelf" following "Put the book on the table". We tested two hypotheses derived from an earlier production study [4]: the "irrelevant intonation" hypothesis vs. the "linguistic intonation" hypothesis. The former predicted that listeners would have a bias towards the contrastive referent regardless of the intonation of *nu*. The latter predicted that listeners would associate an unaccented *nu* with a contrastive referent but an accented *nu* with the referent mentioned in the first instruction.

We observed an initial bias towards the contrastive referent regardless of condition (cf., [6]). Unexpectedly, accentuation on *nu* boosted this initial bias, leading to an immediate increase in fixations to the contrastive referent. This bias was overcome when information on the accentuation of the referent became available. This suggests that accent *nu* may have 'misled' the participants into the interpretation of contrasts in both referent and location. When *nu* was unaccented, there was a steady and significant increase in fixations to the contrastive referent, as predicted. These anticipatory effects of *nu* 'now' stand in contrast to the reported absence of anticipatory effects for a number of discourse markers (Ito & Speer [9]).

These findings accord with the "linguistic intonation" hypothesis and show that listeners make use of prosodic information prior to the entire referential expression to identify the upcoming referent.

5. References

[1] Baayen, R. H.; Piepenbrock, R.; Gulikers, L., 1995. [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

[2] Barr, D.J. (in press). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*

[3] Bates, D. & Sarkar, D. (2007). Lme4: Linear mixed-effects models using S4 classes. R package version 0.9975-13

[4] Braun, B.; Chen, A., 2007. And now for something completely different: Intonation of 'now' and scope ambiguity in English and Dutch. AMLaP. Turku, Finland.

[5] Dahan, D.; Magnuson, J.; Tanenhaus, M., 2001. Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cogn. Psy*, 42, 317-367.

[6] Dahan, D.; Tanenhaus, M.; Chambers, C.G., 2002. Accent and reference resolution in spoken-language comprehension. *J. of Memory and Language* 47, 292-314.

[7] Gussenhoven C. (2004). Transcription of Dutch intonation. In: Sun-Ah Jun (ed) Prosodic Typology and Transcription: A Unified Approach. CUP. pp. 118-145.

[8] Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19, 501-530.

[9] Ito, K. and Speer, S.R. (in press). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*

[10] Matin, E.; K. Shao; Boff, K., 1993. Saccadic overhead: Information processing time with and without saccades. *Perceptual Psychophysics* 53, 372-380.

[11] McQueen, J.M.; Viebahn, M., 2007. Tracking recognition of spoken words by tracking looks to printed words. *The Quarterly J. of Experimental Psychology* 60(5), 661-671.

[12] Weber, A.; Braun, B.; Crocker, M., 2006. Finding referents in time: eye-tracking evidence for the role of contrastive accents. *Language & Speech* 49(3), 367-392.