Prosody and the Resolution of Pronominal Anaphora

Maria Wolters
Institut für Kommunikationsforschung und Phonetik, Universität Bonn
Poppelsdorfer Allee 47, D-53115 Bonn
wolters@ikp.uni-bonn.de

Donna K. Byron
Department of Computer Science
University of Rochester
P.O. Box 270226, Rochester, NY 14627
dbyron@cs.rochester.edu

Abstract

In this paper, we investigate the acoustic prosodic marking of demonstrative and personal pronouns in task-oriented dialog. Although it has been hypothesized that acoustic marking affects pronoun resolution, we find that the prosodic information extracted from the data is not sufficient to predict antecedent type reliably. Inter-speaker variation accounts for much of the prosodic variation that we find in our data. We conclude that prosodic cues should be handled with care in robust, speaker-independent dialog systems.

1 Introduction

Previous work on anaphora resolution has yielded a rich basis of theories and heuristics for finding antecedents. However, most research to date has neglected an important potential cue that is only available in spoken data: prosody. Prosodic marking can be used to change the antecedent of a pronoun, as demonstrated by this classic example from Lakoff (1971) (capitals indicate a pitch accent):

(1) John$_i$ called Jim$_j$ a Republican, then he$_i$ insulted him$_j$.

(2) John$_i$ called Jim$_j$ a Republican, then HE$_j$ insulted HIM$_i$.

But exactly how the antecedent changes due to the prosodic marking on the pronoun, and whether this effect happens consistently, is an open question. If consistent effects do exist, they would be useful for online pronoun interpretation in spoken dialog systems.

Prosodic prominence directs the attention of the listener to what is important for understanding and interpretation. But how should this principle be applied when words that are normally not very prominent, such as pronouns, are accented? More generally, does acoustic marking provide systematic cues to characteristics of antecedents? More specifically, does it imply that the antecedent is "unusual" in some way? These are the two hypotheses we investigate in this paper. Our data consists of 322 pronouns from a large corpus of spontaneous task-oriented dialog, the TRAINS93 corpus (Heeman and Allen, 1995). This corpus allows us to study pronouns as they occur in spontaneous unscripted discourse, and is one of the very few speech corpora to have been annotated with pronoun interpretation information.

The remainder of this paper is structured as follows: In Section 2, we summarize relevant work on pronoun resolution and report on the few proposals for integrating prosody into pronoun resolution algorithms. Next, in Section 3, we present the dialogs used for our study and the attributes available in the annotation data, while Section 4 describes the acoustic measures that were computed automatically from the data. Section 5 explores whether there are systematic correlations between these properties and the acoustic measures fundamental frequency, duration, and intensity. For these measures, we find that most correlations are in fact due to speaker variation, and that speakers differ greatly in their overall prosodic characteristics. Finally, we investigate whether it is possible to use these acoustic features to predict properties of the antecedent using logistic regression. Again, we do not find acoustic features to be reliable predictors for the features of interest. Therefore, we conclude in Section 6 that acoustic measures cannot be used in speaker–independent online anaphora resolution algorithms to predict the features under investigation here.

2 Background and Related Work

There is a rich literature on resolving personal pronouns. Many approaches are based on a notion of attentional focus. Entities in attentional focus are highly salient, and pronouns are assumed to refer to the most salient entity in the discourse (cf. (Brennan et al., 1987; Azzam et al., 1998; Strube, 1998)). Centering (Grosz et al., 1995) is a framework for predicting local attentional focus. It assumes that the most salient entity from sentence S_{n-1} that is realized in sentence S_n is most likely to be pronominalized in S_n. That entity is termed the *Cb* (backward-looking center) of sentence S_n. Finding the preferred ranking criteria is an active area of research. Byron and Stent (1998) adapted this approach, which had previously been applied to text, for spoken dialogs, but with limited success.

In contrast to personal pronouns, demonstratives do not rely on calculations of salience. In fact, Linde (1979) found that while *it* was preferred for entities within the

current local focus, *that* was used for items outside the current focus of attention. Passonneau (1989) showed that personal and demonstrative pronouns are used in contrasting situations: personal pronouns are preferred when both the pronoun and its antecedent are in subject position, while demonstrative pronouns are preferred when either the pronoun or its antecedent is not in subject position. She also found that personal pronouns tend to co-specify with pronouns or base noun phrases; the more clause- or sentence-like the antecedent, the more likely the speaker is to choose a demonstrative pronoun.

Pronoun resolution algorithms tend not to cover demonstratives. Notable exceptions are Webber's model for discourse deixis (Webber, 1991) and the model developed for spoken dialog by Eckert and Strube (1999). This algorithm encompasses both personal and demonstrative pronouns and exploits their contrastive usage patterns, relying on syntactic clues and verb subcategorizations as input. Neither study investigated the influence of prosodic prominence on resolution.

Most previous work on prosody and pronoun resolution has focussed on pitch accents and third person singular pronouns that co-specify with persons. Nakatani (1997) examined the antecedents of personal pronouns in a 20-minute narrative monologue. She found that pronouns tend to be accented if they occur in subject position, and if the backward-looking center (Grosz et al., 1995) was shifted to the referent of that pronoun. She then extended this result to a general theory of the interaction between prominence and discourse structure. Cahn (1995) discusses accented pronouns on the basis of a theory about accentual correlates of salience. Kameyama (1998) interprets a pitch accent on pronouns in the framework of the alternative semantics (Rooth, 1992) theory of focus. She assumes that all potential antecedents are stored in a list. Pronouns are then resolved to the most preferred antecedent on that list which is syntactically and semantically compatible with the pronoun. Preference is modeled by an ordering on the set of antecedents. An accent on the pronoun signals that pronoun resolution should not be based on the default ordering, where the default is computed by a number of interacting syntactic, semantic, pragmatic, and attentional constraints.

Compared to *he* and *she*, *it* and *that* have been somewhat neglected. There are two reasons for this: First, *it* is not considered to be as accentable as *he* and *she* by native speakers of both British and American English, whereas *that* is more likely than *it* to bear a pitch accent. An informal study of the London-Lund corpus of spoken British English (Svartvik, 1990) confirmed that observation. Second, *that* frequently does not have a co-specifying NP antecedent, and most research on co-specification has focussed on pronouns and NPs. Work on accented demonstratives and pronoun resolution is extremely scarce. Pioneering studies were conducted by Fretheim and his collaborators. They tested the effect of accented sentence-initial demonstratives that co-specify with the preceding sentence on the resolution of ambiguous personal pronouns, and found that the pronoun antecedents switched when the demonstrative was accented (Fretheim et al., 1997). However, to our knowledge, there are no studies that compare the co-specification preferences of accented vs. unaccented demonstratives.

3 The Corpus: TRAINS93

Our data is taken from the TRAINS93 corpus of human-human problem solving dialogs in the logistics planning domain. In these dialogs, one participant plays the role of the planning assistant and the other attempts to construct a plan for delivering specified cargo to its destination. We used a subset of 18 TRAINS93 dialogs in which the referent and antecedent of third-person non-gendered pronouns[1] had been annotated in a previous study (Byron and Allen, 1998). In the dialogs used for the present study, 322 pronouns (158 personal and 164 demonstrative) have been annotated. Personal pronouns in the dialogs are *it*, *its*, *itself*, *them*, *they*, *their* and *themselves*. Demonstrative pronouns in the annotation data are *that*, *this*, *these*, *those*. There are five male and 11 female speakers. One female speaker contributed 89 pronouns, two others produced more than 30 each (one female, one male), the rest is divided unevenly among the remaining 13 speakers. The set of dialogs chosen for annotation intentionally included a variety of speakers so that no speaker's idiosyncratic discourse strategies would be prevalent in the resulting data.

Table 1 describes the attributes captured for each pronoun. These features were chosen for the annotation because many previous studies have shown them to be important for pronoun resolution. Features include attributes of the pronoun, its antecedent (the discourse constituent that previously triggered the referent), and its referent (the entity that should be substituted for the pronoun in a semantic representation of the sentence). Cb was annotated using Model3 from (Byron and Stent, 1998) with a linear model of discourse structure. Note that annotated pronouns were not limited to those with NP antecedents, as is the case with most other studies. In addition to NP antecedents, pronouns in this data set could have an antecedent of some other phrase or clause type, or no annotatable antecedent at all. There are two categories of pronouns with no annotatable antecedent. In the simplest case, the pronominal reference is the first mention of the referent in the dialog. That happens when the referent is inferred from the problem solving state. For example, after the utterance `send the engine to Corning and pick up the boxcars`, a new discourse en-

[1] No gendered entities exist in this corpus, so gendered pronouns were not included. All demonstrative pronouns were annotated; however, there were only 5 occurrences of "this" in the selected dialogs, so constrasts between proximal and distal demonstratives could not be studied.

Feature ID	Description	Possible Values
PRONTYPE	Pronoun Type	def = the pronoun is one of {it, its, itself, them, they, their, themselves} dem = the pronoun is one of {that, this, these, those}
PRONSUBJ	Pronoun is subject	Y = pronoun subject of main clause of its utterance N = pronoun not subject of main clause
ANTEFORM	Antecedent form	PRONOUN = antecedent is pronoun NP = antecedent is base noun phrase NON-NP = antecedent is other constituent, at most one utterance long NONE = pronoun is first mention or antecedent length > one utterance
DIST	Distance to antecedent	SAME = antecedent and pronoun in same utterance ADJ = antecedent and pronoun in adjacent utterances REMOTE = antecedent more than one utterance before pronoun
ANTESUBJ	Antecedent is subject	Y = antecedent subject of the main clause of its utterance N = antecedent not subject of a main clause
CB	Backward-looking center	Y = pronoun is Cb of its utterance N = pronoun is not Cb

Table 1: The features available in the annotation data set.

Pronoun	ANTE			ANTESUBJ		DIST		
category	*NP/pron.*	*non-NP*	*none*	*yes*	*no*	*same*	*adj.*	*remote*
personal	**75.9**%	6.3%	17.8 %	37.3%	**62.7**%	29.1%	**33.5**%	20.2%
demonstrative	28.0%	**36.0**%	36.0%	14.0%	**86.0**%	18.9%	29.9%	15.2%
total	**51.6**%	21.4%	27.0%	25.5%	**74.5**%	23.9%	**31.7**%	17.7%

Table 2: Typical properties of antecedents for personal and demonstrative pronouns in the corpus. All percentages are given relative to the total number of pronouns in that category and rounded. Boldface: most frequent antecedent property.

tity, the train composed of the engine and boxcars, is available for anaphoric reference. In the more subtle case, the entity was built from a stretch of discourse longer than one utterance. In an effort to achieve an acceptable level of inter-annotator agreement for the annotation, the maximum size for a constituent to serve as an antecedent was defined to be one utterance. Discourse entities that are built from longer stretches of text include objects such as the entire plan or the discourse itself, and such items are less reliable to annotate.

Taking the annotated dialogs as a whole, 21.4% of all pronouns have a non-NP antecedent, and 27% do not have an annotatable antecedent at all. Table 2 shows that the default antecedents of personal and demonstrative pronouns follow the predictions of Schiffman (1985). The antecedent of personal pronouns is most likely itself to be a pronoun or a full NP, while demonstratives are most likely to have no antecedent, or if there is one, it is most likely to be a non-NP. The main role of prosodic information is to help pronoun resolution algorithms identify cases where these default predictions are false.

4 Acoustic Prosodic Cues

Our selection of acoustic measures covers three classic components of prosody: fundamental frequency (F0), duration, and intensity (Lehiste, 1970). The relationship between those cues and prosodic prominence has been demonstrated by e.g. (Fant and Kruckenberg, 1989; Heuft, 1999). The main correlate of English stress is F0, the second most important is duration, and the least important is intensity (Lehiste, 1970). Therefore, we will pay more attention to F0 measures. Although experimental results indicate that F0 cues of prominence can depend on the shape of the F0 contour of the utterance (c.f. (Gussenhoven et al., 1997)), we do not control for such interactions. Instead, we restrict ourselves to cues that are easy to compute from limited data, so that a running spoken dialogue system might be able to compute them in real time.

4.1 Acoustic Measures

Duration: For duration, we found that the logarithmic duration values are normally distributed, both pooled over all speakers and for those speakers with more than 20 pronouns. Logarithmic duration is also the target variable of many duration models such as that of (van Santen, 1992). We assume that speaker-related variation is covered by the variance of this normal distribution; we can control for speaker effects by including a SPEAKER factor in our models.

F0 variables: F0 was computed using the Entropic ESPS Waves tool `get_f0` with standard settings and a frame rate of 10 ms. All F0 values were transformed into the log-domain and then pooled into mean, minimum, and maximum F0 values for each word and each utterance. This log domain is well motivated psychoacoustically (Zwicker and Fastl, 1990). F0 range was computed on the values in the log-domain. We assume that the logarithm of F0 has a normal distribution. Therefore, we

can normalize for speaker-dependent differences in pitch range by using z-scores, and we can use standard statistical analysis methods such as ANOVA.

Intensity: Intensity is measured as the root-mean-square (RMS) of signal amplitudes. We measure RMS relative to a baseline as given by the formula $\log(\mathrm{RMS}/\mathrm{RMS}_{\mathrm{baseline}})$. The baseline RMS was computed on the basis of a simple pause detection algorithm, which takes the first maximum in the amplitude histogram to be the average amplitude of background noise. The baseline RMS was slightly above that value.

4.2 Inter-Speaker Differences

Since we need to pool data from many different speakers, we need to control for inter-speaker differences. The number of pronouns we have from each speaker varies between 1 for speaker GD and 86 for speaker CK. Speakers PH, male, and CK, female, are the only ones to have produced more than 15 personal pronouns and 15 demonstratives. In order to test whether the SPEAKER factor affects the choice between personal pronouns and demonstratives, we fitted a logistic regression model with the target variable PRONTYPE (personal or demonstrative) and the predictors ANTE, ANTESUBJ, DIST, REFCAT, CB and SPEAKER (in this sequence). REFCAT is an additional variable that describes the semantic category of a pronoun's referent (eg. domain objects vs. abstract entities). Even though SPEAKER is the last factor in the model, an analysis of deviance shows a significant influence ($p<0.005$,$F=2.51$,$df=13$). A possible explanation for this is that some speakers prefer to use demonstratives in contexts where others would choose a personal pronoun, and vice versa, or perhaps the SPEAKER variable mediates the influence of a far more complex factor such as problem solving strategy. Resolving this question is beyond the scope of this paper.

On the basis of F0, we can establish four groups of speakers: The first group consists of male speakers with a low mean F0 and a low F0 range. In the next group, we find both male and female speakers with a low mean F0, but a far higher range. Speaker PH belongs to this second group. Interestingly, for these speakers, the mean F0 on pronouns is lower than for those of the first group. Groups 3 and 4 consist entirely of female speakers, with group 3 using a lower range than group 4. Speaker CK belongs to group 4.

5 Exploring Prominent Pronouns

If data about prosodic prominence is to be useful for pronoun resolution, then there must be prosodic cues that carry information about properties of the antecedent. In this section, we investigate if there are such cues for the properties that we have available in the annotation data, defined in Table 1. More specifically, we hypothesize that prosodic cues will be used if the antecedent is somewhat unusual. For example, the results of Linde and Passonneau would lead us to expect that personal pronouns with non-NP antecedents and demonstratives with NP and pronoun antecedents will be marked. Since the antecedents of pronouns tend to occur no more than 1-2 clauses ago, we would also expect pronouns with more remote antecedents to be marked. A first qualitative look at the data suggets that even if such these tendencies are present in the data, they might not turn out to be significant. For example, in Figure 1, the means of lzmeanf0 behave roughly as predicted, but the variation is so large that these differences might well be due to chance.

Property	**df**	**Data Set**			
		all	*pers.*	*dem.*	*CK*
ANTEFORM	3	range	none	none	none
DIST	3	none	none	none	none
ANTESUBJ	2	dur	dur, mean	none	pers.: energy range

Table 3: Significant Influences of Antecedent Properties ($p<0.05$) on Prosodic Cues. mean=z-score mean F0, range=range of z-score F0, dur=logarithmic duration, dem=demonstratives, pers=personal pronouns

5.1 Correlations between Measures and Properties

Next, we examine whether the measures defined in Section 4 correlate with any particular properties of the antecedent. More precisely, if a property is cued by some aspect of prosody (either duration, F0, or intensity), then the prosody of a pronoun depends to a certain degree on its antecedent. In a statistical analysis, we should find a significant effect of the relevant antecedent property on the prosodic measure. We selected ANOVA as our analysis method, because our prosodic target variables appear to have a normal distribution. For each of the antecedent features defined above, we examined its influence on mean F0 (lmeanf0), the z-score of mean F0 (lzmeanf0), the z-score of F0 range (lzrgf0), logarithmic duration (dur), and normalized energy (energy). In addition, we added the factors, PRONTYPE and SPEAKER.

Results: The results are summarized in Table 3. For lzmeanf0 and energy, the influence of SPEAKER is always considerable. There are also consistent effects of the syntactic position of a pronoun: In general, demonstratives are shorter in subject position, and for CK, mean F0 on personal pronouns in subject position is higher than on non-subject ones (228 Hz vs. 190 Hz). But when we turn to the factors that interest us most, properties of the antecedent, we cannot find any consistent correlates, although in almost every data set, there are some prosodic cues to ANTESUBJ for personal pronouns. But what these cues are may well depend on the speaker, as the results for CK show. Her pitch range on pronouns with a subject antecedent is double the range on pronouns with an antecedent in non-subject position.

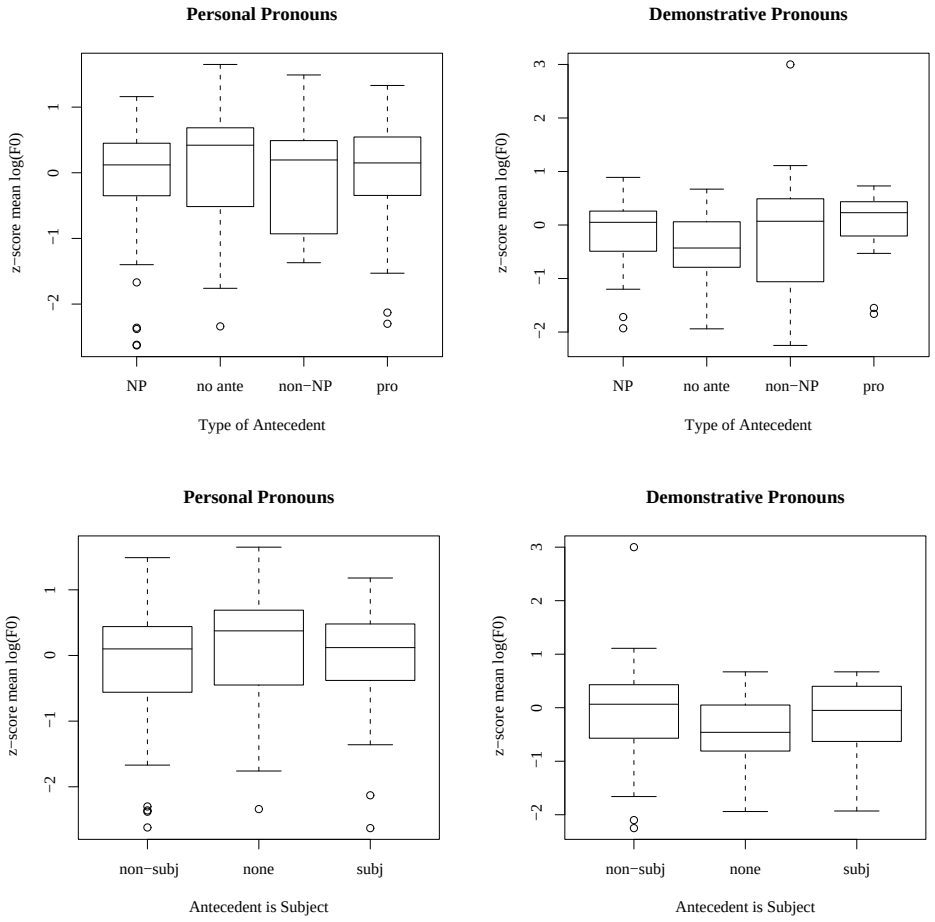

Figure 1: Distribution of z-score of mean F0 for different values of ANTEFORM and ANTESUBJ

Pronouns with subject antecedents are also considerably louder. All in all, antecedent properties can only account for a very small percentage of the variation in these prosodic cues. Therefore, we should not expect the prosodic cues to be stable, robust indicators for predicting antecedent properties in spoken dialog systems.

5.2 Inter-Speaker Variation

We have seen that inter-speaker differences explain much of the variation in the prosodic measures. Table 4 gives an idea of the size and direction of these differences.

On the complete data set, we find that personal pronouns are shorter than demonstratives, they have a lower intensity and show a higher average F0 (Table 4). A closer examination reveals considerable inter-speaker variation in the data, illustrated in Table 4. CK is fairly prototypical. PH barely shows the difference in F0, and for MF, the difference in intensity is actually reversed. MF also has rather short demonstratives. Such speaker-specific variation cannot be eliminated by normalization. It has to be controlled for in the statistical tests. Discovering types of speakers is difficult – two of the 15 speakers, CK, and PH, contribute 48% of all pronouns.

5.3 Predicting Properties of the Antecedent

Finally, we examine how much information prosodic cues yield about the antecedent. For this purpose, we set up a prediction task not unlike one that an actual NLU system faces. The input variables are the prosodic properties of the pronoun, whether the pronoun is personal or demonstrative (PRONTYPE), whether it is the subject (PRONSUBJ), and whether it is sentence-initial (PRONINIT). From this, we now have to deduce properties of the antecedent: syntactic role (ANTESUBJ), form (ANTEFORM), and distance (DIST). For prediction, we used logistic regression (Agresti, 1990). This has two advantages: not only can we compare how well the different regression models fit the data, we can also re-analyze the fitted model to determine which factors have a significant influence on classification accuracy.

First, we construct a model on the basis of PRONTYPE, PRONSUBJ, and PRONINIT. Then, we construct a model with these three factors plus SPEAKER. Finally, we train a model with PRONTYPE,

Speaker	mean F0			z-score mean		duration		intensity	
	disc.	*pers.*	*dem.*	*pers.*	*dem*	*pers.*	*dem*	*pers.*	*dem*
all	156 Hz	157 Hz	142 Hz	-0.04	-0.24	161 ms	206 ms	2.36	2.38
CK	188 Hz	208 Hz	187 Hz	0.31	0.00	151 ms	193 ms	2.51	2.54
PH	126 Hz	109 Hz	110 Hz	*-0.43*	-0.47	179 ms	252 ms	2.57	2.84
MF	166 Hz	184 Hz	182 Hz	0.32	0.26	*166 ms*	164 ms	*2.69*	2.40

Table 4: Inter-speaker variation in prosody. *disc.*: complete discourse. All speakers: 322 pronouns, CK: 41 personal, 45 demonstrative, PH: 18 personal, 24 demonstrative, MF: 7 personal, 8 demonstrative

PRONSUBJ, PRONINIT, SPEAKER and one of the three measures lzmeanf0,dur,energy. The models are trained to predict whether there is an antecedent (task noAnte), whether the antecedent is a non-NP (task nonNP), whether the antecedent is remote (task remote), whether the antecedent is in subject position (task sjante), and whether the antecedent is the current Cb (task cb). All models are computed over the full data set, because the data set for speaker CK is not sufficient for estimating the regression coefficients. The models are then compared to see which step yielded a significant improvement: adding SPEAKER or adding the prosodic variable after we have accounted for SPEAKER variation.

Results: The results are summarized in Table 5. On all tasks except remote, PRONTYPE and PRONSUBJ performed well. Both features have already been shown to be reliable cues for pronoun resolution (c.f. Section 2). On task cb, only PRONTYPE can explain a significant amount of variation. Models which include a speaker factor almost always fare better. In models without speaker information, F0-related measures yield a larger reduction in deviance than the duration measure. The reason for this is that the F0 measures preserve some information about the different speaker strategies. Once SPEAKER has been included as well, only dur leads to significant improvements on task nonNP ($p<0.05$). Both demonstratives and personal pronouns are shorter when the antecedent is a non-NP.

6 Conclusion and Outlook

In this paper, we examined patterns of acoustic prosodic highlighting of personal and demonstrative pronouns in a corpus of task-oriented spontaneous dialog. To our knowledge, this is the first comparative study of this kind. We used a straightforward, theory-neutral operationalization of "prosodic highlighting" that does not depend on complex algorithms for F0 stylization or (focal) accent detection and is thus very easy to incorporate into any real-time spoken dialog system. We chose a spoken dialog corpus that includes demonstrative pronouns because demonstratives are both a prominent feature of problem-solving dialogs and a sorely neglected field of study. In particular, we asked two questions:

Do Speakers Signal Antecedent Properties Acoustically? Based on our data, the answer to this question is: If they do, they do it in a highly idiosyncratic way. We cannot posit any safe generalizations over several speakers, and from the perspective of an NLP application, such generalizations might even be dangerous. In order to evaluate the impact of speaker strategies on the resolution of pronouns, we need more data – 150 to 200 pronouns from 4-5 speakers each. Collecting this amount of data in a dedicated corpus is inefficient. Therefore, further acoustic investigations do not make much sense at this point; rather, the data should be examined carefully for tendencies which can form the basis for dedicated production and perception experiments which are explicitly designed for uncovering inter-speaker variation.

Are Acoustic Features Useful for Pronoun Resolution? The answer is: probably not. At least for this corpus, we were not able to determine any numerical heuristics that could be utilized to aid pronoun resolution. The logistic regression experiments show that on a speaker-independent basis, logarithmic duration might well be a reliable cue to certain aspects of a pronoun's antecedent. In order to incorporate prosodic cues into an actual algorithm, we will need more training material and a principled evaluation procedure. We will also need to take into account other influences, such as dialog acts and dialog structure.

Acknowledgements. We would like to thank the three anonymous reviewers, Rebecca Passonneau, Lucien Galescu, James Allan, Michael Strube, Dietmar Lancé and Wolfgang Hess for their comments on earlier versions of this work. Donna K. Byron was funded by ONR research grant N00014-95-1-1088 and Columbia University/NSF research grant OPG:1307. For all statistical analyses, we used R (Ihaka and Gentleman, 1996).

References

A. Agresti. 1990. *Categorical Data Analysis*. John Wiley.

S. Azzam, K. Humphreys, and R. Gaizauskas 1998. Extending a Simple Coreference Algorithm with a Focusing Mechanism. In *New Approaches to Discourse Anaphora: Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*, pages 15–27.

S. Brennan, M. Friedman, and C. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL '87)*, pages 155–162.

Task	significant influence
nonNP	PRONTYPE, PRONSUBJ, PRONINIT, dur
noAnte	PRONTYPE, PRONSUBJ, PRONINIT, SPEAKER
remote	none
sjante	PRONTYPE, PRONSUBJ
cb	PRONTYPE, SPEAKER

Table 5: Performance of Regression Models on Tasks. Listed are factors which improve performance significantly ($p < 0.05$)

D. Byron and J. Allen. 1998. Resolving demonstrative pronouns in the TRAINS93 corpus. In *New Approaches to Discourse Anaphora: Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*, pages 68 – 81.

D. Byron and A. Stent. 1998. A preliminary model of centering in dialog. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL '98)*.

J. Cahn. 1995. The effect of pitch accenting on pronoun referent resolution. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 290–292.

M. Eckert and M. Strube. 1999. Resolving discourse deictic anaphora in dialogs. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*.

G. Fant and A. Kruckenberg. 1989. Preliminaries to the study of Swedish prose reading and reading style. *KTH Speech Transmission Laboratory Quarterly Progress and Status Report*, 2:1–83.

T. Fretheim, W. van Dommelen, and K. Borthen. 1997. Linguistic constraints on relevance in reference resolution. In K. Singer, R. Eggert, and G. Anderson, editors, *CLS*, volume 33, pages 99–113.

B. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

C. Gussenhoven, B.H. Repp, A. Rietveld, H. Rump and J. Terken. 1997. The perceptual prominence of fundamental frequency peaks. *J. Acoust. Soc. Amer.*, 102:3009–3022.

P. Heeman and J. Allen. 1995. The Trains Spoken Dialog Corpus. CD-ROM, Linguistic Data Consortium.

B. Heuft. 1999. *Eine prominenzbasierte Methode zur Prosodieanalyse und -synthese*. Peter Lang, Frankfurt.

R. Ihaka and R. Gentleman (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

M. Kameyama. 1998. Stressed Pronouns. In P. Bosch, R. van Sandt, editors, *The Focus Book*, pages 89–112. Oxford University Press, Oxford.

G. Lakoff. 1971. Presuppositions and relative well-formedness. In *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, and Psychology*, pages 329–340. Cambridge University Press.

I. Lehiste. 1970. *Suprasegmentals*. MIT Press, Cambridge, Mass.

C. Linde. 1979. Focus of attention and the choice of pronouns in discourse. In Talmy Givon, editor, *Syntax and Semantics 12: Discourse and Syntax*, New York. Academic Press.

C. Nakatani. 1997. The Computational Processing of Intonational Prominence: A Functional Prosody Perspective. Ph.D. thesis, Harvard University.

R. Passonneau. 1989. Getting at discourse referents. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL '89)*, pages 51–59.

M. Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1:75–112.

R. Schiffman (Passonneau). 1985. *Discourse constraints on 'it' and 'that': A study of language use in career-counseling interviews*. Ph.D. thesis, University of Chicago.

J. Svartvik, editor. 1990. *The London Corpus of Spoken English: Description and Research.* Lund University Press, Lund.

M. Strube 1998. Never look back: An alternative to centering. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL '98)*, pages 1251–1257.

J. van Santen 1992 Contextual effects on vowel duration. *Speech Communication*, 11:513–546.

B. Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6:107-135.

E. Zwicker and H. Fastl 1990. Psychoacoustics. Springer, Berlin.