Hand Gestures in Disambiguating Types of *You* Expressions in Multiparty Meetings

Tyler Baldwin
Department of Computer
Science and Engineering
Michigan State University
East Lansing, MI 48824
baldwin96@cse.msu.edu

Joyce Y. Chai
Department of Computer
Science and Engineering
Michigan State University
East Lansing, MI 48824
jchai@cse.msu.edu

Katrin Kirchhoff
Department of Electrical
Engineering
University of Washington
Seattle, WA, USA
katrin@ee.washington.edu

Abstract

The second person pronoun *you* serves different functions in English. Each of these different types often corresponds to a different term when translated into another language. Correctly identifying different types of *you* can be beneficial to machine translation systems. To address this issue, we investigate disambiguation of different types of *you* occurrences in multiparty meetings with a new focus on the role of hand gesture. Our empirical results have shown that incorporation of gesture improves performance on differentiating between the generic use of *you* (e.g., refer to people in general) and the referential use of *you* (e.g., refer to a specific person or a group of people). Incorporation of gesture can also compensate for limitations in automated language processing (e.g., reliable recognition of dialogue acts) and achieve comparable results.

1 Introduction

The second person pronoun *you* is one of the most prevalent words in conversation and it serves several different functions (Meyers, 1990). For example, it can be used to refer to a single addressee (i.e., the *singular* case) or multiple addressees (i.e., the *plural* case). It can also be used to represent people in general (i.e., the *generic* case) or be used idiomatically in the phrase "you know".

For machine translation systems, these different types of *you* often correspond to different translations in another language. For example, in German, there are different second-person pronouns for singular vs. plural *you* (viz. *du* vs. *ihr*); in addition there are different forms for formal vs. informal forms of address (*du* vs. *Sie*) and for the generic use (*man*). The following examples

demonstrate different translations of *you* from English (EN) into German (DE):

- Generic *you*
  EN: Sometimes **you** have meetings where the decision is already taken.
  DE: Manchmal hat **man** Meetings wo die Entscheidung schon gefallen ist.
- Singular *you*:
  EN: Do **you** want an extra piece of paper?
  DE: Möchtest **du** noch ein Blatt Papier?
- Plural *you:*
  EN: Hope **you** are all happy!
  DE: Ich hoffe, **ihr** seid alle zufrieden!

These examples show that correctly identifying different types of *You* plays an important role in the correct translation of *you* in different context.

To address this issue, this paper investigates the role of hand gestures in disambiguating different usages of *you* in multiparty meetings. Although identification of *you* type has been investigated before in the context of addressee identification (Gupta et al., 2007b; Gupta et al., 2007a; Frampton et al., 2009; Purver et al., 2009), our work here focuses on two new angles. First, because of our different application on machine translation, rather than processing *you* at an utterance level to identify addressee, our work here concerns each occurrence of *you* within each utterance. Second and more importantly, our work investigates the role of corresponding hand gestures in the disambiguation process. This aspect has not been examined in previous work.

When several speakers are conversing in a situated environment, they often overtly gesture at one another to help manage turn order or explicitly direct a statement toward a particular participant (McNeill, 1992). For example, consider the following snippet from a multiparty meeting:

A: "Why is that?"
B: "Because, um, based on what ev-

erybody's saying, right, [*gestures at Speaker D*] <u>you</u> want something simple. <u>You</u> [*gestures at Speaker C*]want basic stuff and [*gestures at Speaker A*] <u>you</u> want something that is easy to use. Speech recognition might not be the simplest thing."

The use of gesture in this example indicates that each instance of the pronoun *you* is intended to be referential, and gives some indication of the indented addressee. Without the aid of gesture, it would be difficult even for a human listener to be able to interpret each instance correctly.

Therefore, we conducted an empirical study on several meeting segments from the AMI meeting corpus. We formulated our problem as a classification problem for each occurrence of *you*, whether it is a *generic*, *singular*, or *plural* type. We combined gesture features with several linguistic and discourse features identified by previous work and evaluated the role of gesture in two different settings: (1) a two stage classification that first differentiates the *generic* type from the *referential* type and then within the *referential* type distinguishes *singular* and *plural* usages; (2) a three way classification between *generic*, *singular*, or *plural* types. Our empirical results have shown that incorporation of gesture improves performance on differentiating between the *generic* and the *referential* type. Incorporation of gesture can also compensate for limitations in automated language processing (e.g., reliable recognition of dialogue acts) and achieve comparable results. These findings have important implications for machine translation of *you* expressions from multiparty meetings.

2 Related Work

Psychological research on gesture usage in human-human dialogues has shown that speakers gesture for a variety of reasons. While speakers often gesture to highlight objects related to the core conversation topic (Kendon, 1980), they also gesture for dialogue management purposes (Bavelas et al., 1995). While not all of the gestures produced relate directly to the resolution of the word *you*, many of them give insight into which participant is being addressed, which has a close correlation with *you* resolution. Our investigation here is closely related to two areas of previous work: addressee identification based on *you* and the use of gestures in coreference resolution.

Addressee Identification. Disambiguation of *you* type in the context of addressee identification has been examined in several papers in recent years. Gupta et. al. (2007b) examined two-party dialogues from the Switchboard corpus. They modeled the problem as a binary classification problem of differentiating between generic and referential usages (referential usages include the singular and plural types). This work has identified several important linguistic and discourse features for this task (which was used and extended in later work and our work here). Later work by the same group (Gupta et al., 2007a) examined the same problem on multiparty dialogue data. They made adjustments to their previous methods by removing some oracle features from annotation and applying simpler and more realistic features. A recent work (Frampton et al., 2009) has examined both the generic vs. referential and singular vs. plural classification tasks. A main difference is that this work incorporated gaze feature information in both classification tasks (gaze features are commonly used in addressee identification). More recent work (Purver et al., 2009) discovered that large gains in performance can be achieved by including n-gram based features. However, they found that many of the most important n-gram features were topic specific, and thus required training data consisting of meetings about the same topic.

Gestures in Coreference Resolution. Eisenstein and Davis (2006; 2007) examined coreference resolution on a corpus of speaker-listener pairs in which the speaker had to describe the workings of a mechanical device to the listener, with the help of visual aids. In this gesture heavy dataset, they found gesture data to be helpful in resolving references. In our previous work (2009), we examined gestures for the identification of coreference on multparty meeting data. We found that gestures only provided limited help in the coreference identification task. Given the nature of the meetings under investigation, although gestures have not been shown to be effective in general, they are potentially helpful in recognizing whether two linguistic expressions refer to a same participant.

Compared to these two areas of earlier work, our investigation here has two unique aspects. First, as mentioned earlier, previous work on addressee identification focused the problem at the

utterance level. Because the goal was to find the addressee of an utterance, the assumption was that all instances of *you* in an utterance were of the same type. However, since several instances of *you* in the same utterance may translate differently, we instead examine the classification task at the instance level. Second, our work here specifically investigates the role of gestures in disambiguation of different types of *you*. This aspect has not been examined in previous work.

3   Data

The dataset used in our investigation was the AMI meeting corpus (Popescu-Belis and Estrella, 2007), the same corpus used in previous work (Gupta et al., 2007a; Frampton et al., 2009; Purver et al., 2009; Baldwin et al., 2009). The AMI meeting corpus is a large publicly available corpus of multiparty design meetings. AMI meeting annotations contain manual speech transcriptions, as well as annotations of several additional modalities, such as focus of attention and head and hand gesture.

For this work, six AMI meeting segments (IS1008a, IS1008b, IS1008c, IS1008d, ES2008a, TS3005a) were used. These instances were chosen because they contained manual annotations of hand gesture data, which was not available for all AMI meeting segments. These six meeting segments were from AMI "scenario" meetings, in which meeting participants had a specific task of designing a hypothetical remote control.

All instances of the word *you* and its variants were manually annotated as either generic, singular, or plural. This produced a small dataset of 533 instances. Agreement between two human annotators was high ($\kappa = 0.9$). The distribution of *you* types is shown in Figure 1. The most prevalent type in our data set was the generic type, which accounted for 47% of all instances of *you* present. Of the two referential types, the singular type accounted for about 60% of the referential instances.

A total of 508 gestures are present in our data set. Table 1 shows the distribution of gestures. As shown, "non-communicative gestures", make up nearly half (46%) of the gestures produced. These are gestures that are produced without an overt communicative intent, such as idly tapping on the table. The other main categorization of gestures is "communicative gestures", which accounts for 45% of all gestures produced and is

made up of the "pointing at participants", "pointing at objects", "interact with object", and "other communicative" gesture types from Table 1. A total of 17% of the gestures produced were pointing gestures that pointed to people, a type of gesture that would likely be helpful for *you* type identification. A small percentage of the gestures produced were not recorded by the meeting recording cameras (i.e., off camera), and thus are of unknown type.

4   Methodology

Our general methodology followed previous work and formulated this problem as a classification problem. We evaluated how gesture data may help *you* type identification using two different approaches: (1) two stage binary classification, and (2) a single three class classification problem. In two stage binary classification, we first attempt to distinguish between instances of *you* that are generic and those that are referential. We then take those cases that are referential and attempt to subdivide them into instances that are intended to refer to a single person and those that refer to several people.

Our feature set includes features used by Gupta et. al. (2007a) (Hereafter referred to as Gupta) and Frampton et. al. (2009) (Hereafter Frampton), as well as new features incorporating gestures. We summarize these features as follows.

Sentential Features. We used several sentential features to capture important phrase patterns. Most of our sentential features were drawn from Gupta (2007a). These features captured the patterns "you guys", "you know", "do you" (and similar variants), "which you" (and variants), "if you", and "you hear" (and variants). Another sentential feature captured the number of times the word *you* appeared in the sentence. Additionally, other features captured sentence patterns not related to *you*, such as the presence of the words "I" and "we".

A few other sentential features were drawn from Frampton et. al. (2009). These include the pattern "<auxiliary> you" (a more general version of the "do you" feature) and a count of the number of total words in the utterances.

Part-of-Speech Features.   Several features based on automatic part-of-speech tagging of the sentence containing *you* were used. Quality of automatic tagging was not assessed. From the tagged results, we extracted 5 features based on sentence
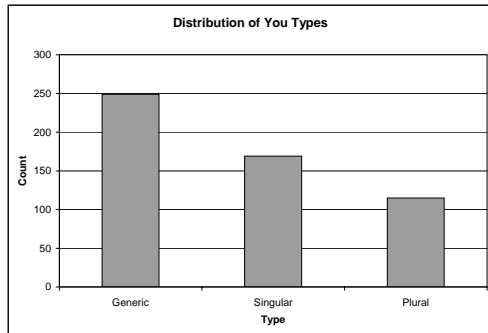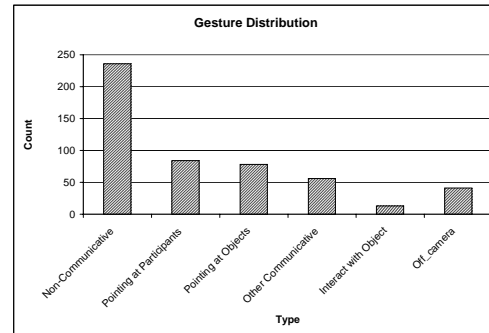
(a) Distribution of *You* types   (b) Distribution of gesture types

Figure 1: Data distributions

and tag patterns: whether or not the sentence that contained *you* also contained *I*, or *we* followed by a verb tag (3 separate features), and whether or not the sentence contains a comparative JJR (adjective) tag. All of these features were adapted from Gupta (2007a).

Dialog Act Features. We used the manually annotated dialogue act tags provided by the AMI corpus to produce our dialogue act features. Three dialogue act features were used: the dialogue act tag of the current sentence, the previous sentence, and the sentence prior to that. Dialog act tags were incorporated into the feature set in one of two different ways: 1) using the full tag set provided by the AMI corpus, and 2) using a binary feature recording if the dialogue act tag was of the elicit type. The latter way of dialogue act incorporation represents a simpler and more realistic treatment of dialogue acts.

Question Mark Feature. The question mark feature captures whether or not the current sentence ends in a question mark. This feature captures similar information to the elicit dialogue act tag and was used in Gupta as an automatically extractable replacement to the manually extracted dialogue act tags (2007a).

Backward Looking/Forward Looking Features. Several features adapted from Frampton et. al. (2009) used information about previous and next sentences and speakers. These features connected the current utterance with previous utterances by the other participants in the room. For each listener, a feature was recorded that indicated how many sentences elapsed between the current sentence and the last/next time the person spoke.

Additionally, two features captured the number of speakers in the previous and next five sentences.

Gesture Features. Several different features were used to capture gesture information. Three types of gesture data were considered: all produced gestures, only those gestures that were manually annotated as being communicative, and only those gestures that were manually annotated as pointing towards another meeting participant. For each of these types, one gesture feature captures the total number of gestures that co-occur with the current sentence, while another feature records only whether or not a gesture co-occurs with the utterance of *you*. Since previous work (Kendon, 1980) has indicated that gesture production tends to precede the onset of the expression, gestures were considered to have co-occurred with instances if they directly overlapped with them or preceded them by a short window of 2.5 seconds.

Note that in this investigation, we used annotated gestures provided by the AMI corpus. Although automated extraction of reliable gesture features can be challenging and should be pursued in the future, the use of manual annotation allows us to focus on our current goal, which is to understand whether and to what degree hand gestures may help disambiguation of *you* Type.

It is also important to note that although previous work (Purver et al., 2009) showed that n-gram features produced large performance gains, these features were heavily topic dependent. The AMI meeting corpus provides several meetings on exactly the same topic, which allowed the n-gram features to learn topic-specific words such as *button, channel, and volume*. However, as real world

	Accuracy
Majority Class Baseline	53.3%
Gupta automatic	70.7%
Gupta manual	74.7%
Gupta + Frampton automatic	73.2%
Gupta + Frampton manual	74.3%
All (+ gesture)	**79.0%**

Table 1: Accuracy values for Generic vs. Referential Classification

	Accuracy
Majority Class Baseline	59.5%
Gupta automatic	72.2%
Gupta manual	73.6%
Gupta + Frampton automatic	73.2%
Gupta + Frampton manual	72.5%
All (+ gesture)	**74.6%**

Table 2: Accuracy values for Singular vs. Plural Classification

meetings occur with a wider range of goals and topics, we would like to build a topic and domain independent model that does not require a corpus of topic specific training data. As such, we have excluded n-gram features from our study.

Additionally, we have not implemented gaze features. Although previous work (Frampton et al., 2009) showed that these features were able to improve performance, we decided to focus solely on gesture to the exclusion of other non-speech modalities. However, we are currently in the process of evaluating the overlap between gesture and gaze feature coverage.

5 Results

Due to the small number of meeting segments in our data, leave-one-out cross validation was preformed for evaluation. Since a primary focus of this paper is to understand whether and to what degree gesture is able to aid in the *you* type identification task, experiments were run using a decision tree classifier due to its simplicity and transparency [1].

5.1 Two Stage Classification

We first evaluated the role of gesture via two stage binary classification. That is, we performed two binary classification tasks, first differentiating between generic and referential instances, and then further dividing the referential instances into the singular and plural types. This provides a more detailed analysis of where gesture may be helpful.

Results for the generic vs. referential and singular vs. plural binary classification tasks are shown in Table 1 and Table 2, respectively. Tables 1 and 2 present several different configurations. The

---

[1] In order to get a more direct comparison to previous work (Gupta et al., 2007a; Frampton et al., 2009), we also experimented with classification via a bayesian network. We found that the overall results were comparable to those obtained with the decision tree.

"Gupta" feature configurations consist of all features used by Gupta et. al. (2007a). These include all part-of-speech features, all dialogue act features, the question mark feature, and all sentential features except the "<auxiliary> you" feature and the word count feature. Results from two types of processing are presented: automatic and manual.

- *Automatic feature extraction (automatic)* - The automatic configurations consist of only features that were automatically extracted from the text. This includes all of the features we examined except for the dialogue act and gesture features. These features are extracted from meeting transcriptions.

- *Manual feature extraction (manual)* - Manual configurations apply manual annotations of dialogue acts and gestures together with the automatically extracted features.

The Frampton configurations add the additional sentential features as well as the backward-looking and forward-looking features. As before, results are presented for a manual and an automatic run. The final configuration ("All") includes the entire feature set with the addition of gesture features. The *All* configuration is the only configuration that includes gesture features.

Although they are not directly comparable, the results for generic vs. referential classification shown in Table 1 appear consistent with those reported by Gupta (2007a). Adding additional features from Frampton et. al. did not produce an overall increase in performance when dialogue act features were present. Including gesture features leads to a significant increase in performance (McNemar Test, $p < 0.01$), an absolute increase of 4.3% over the best performing feature set that does not include gesture. This result seems to confirm our hypothesis that, because gestures are likely

	Accuracy
Majority Class Baseline	46.7%
Gupta automatic	61.5%
Gupta manual	66.2%
Gupta + Frampton automatic	63.6%
Gupta + Frampton manual	70.2%
All (+ gesture)	**70.4%**

Table 3: Accuracy values for several different feature configurations on the three class classification problem.

to accompany referential instances of *you* but not generic instances, gesture information is able to help differentiate between the two. Manual inspection of the decision tree produced indicates that gesture features were among the most discriminative features.

The results on the singular vs. plural task shown in Table 2 are less clear. Although (Gupta et al., 2007a) did not report results on singular vs. plural classification, their feature set produced reasonable classification accuracy of 73.6%. Including gesture and other features did not produce a statistically significant improvement in the overall accuracy. This suggests that while gesture is helpful for predicting referentiality, it does not appear to be a reliable predictor of whether an instance of *you* is singular or plural. Inspection on the decision tree confirms that gesture features were not seen to be highly discriminative.

5.2 Three Class Classification

The results presented for singular vs. plural classification are based on performance on the subset of *you* instances that are referential, which assumes that we are able to filter out generic references with 100% accuracy. While this gives us an evaluation of how well the singular vs. plural task can be performed without the generic references presenting a confounding factor, it presents unrealistic performance for a real system. To account for this, we present results on a three class problem of determining whether an instance of *you* is generic, singular, or plural. The results are shown in Table 3. A simple majority class classifier yields accuracy of 46.7% (In our data, the generic class was the majority class).

As we can see from Table 3, adding additional features gives improved performance over the original implementation by Gupta et. al., re-

sulting in an overall accuracy of about 70%. We also observed that the dialogue act features were important; manual configurations produced absolute gains of about 7% accuracy over fully automatic configurations. The gesture feature, however, did not provide a significant increase in performance over the same feature set without gesture information.

Table 4 shows the precision, recall, and F-measure values for each *you* type for several different configurations. As shown, the generic class proved to be the easiest for the classifiers to identify. This is not suprising, as not only are generic instance our majority class, but many of the features used were originally tailored towards the two class problem of differentiating generic instances from the other classes. The performance on the plural and singular classes is comparable to one another when the basic feature set is used. However, as more features are added, the performance on the singular class increases while the performance on the plural class does not. This seems to suggest that future work should attempt to include more features that are indicative of plural instances.

When manual dialogue acts are applied, it appears incorporation of gestures does not lead to any overall performance improvement (as shown in Table 3). One possible explanation is that gesture features as they are incorporated here do provide some disambiguating information (as shown in the two stage classification), but this information is subsumed by other features, such as dialogue acts. To test this hypothesis, we ran an experiment with a feature set that contained all features except dialogue act features. That is, a feature set that contains all of the automatic features, as well as gesture features. Results are shown in Table 5.

Our "automatic + gesture" feature configuration produced accuracy of 66.2%. When compared to the same feature set without gesture features (the "Gupta + Frampton automatic" row in Table 3) we see a statistically significant ($p < 0.01$) absolute accuracy improvement of about 2.6%. This seems to suggest that gesture features are providing some small amount of relevant information that is not captured by our automatically extractable features.

Up until this point we have incorporated dialogue acts using the full set of dialogue act tags provided by the AMI corpus. As we have men-

		Precision	Recall	F-Measure
Gupta automatic	Plural	0.553	0.548	0.550
	Singular	0.657	0.408	0.504
	Generic	0.624	0.787	0.696
Gupta manual	Plural	0.536	0.513	0.524
	Singular	0.675	0.503	0.576
	Generic	0.704	0.839	0.766
All (+ gesture)	Plural	0.542	0.565	**0.553**
	Singular	0.745	0.604	**0.667**
	Generic	0.754	0.835	**0.792**

Table 4: Precision, recall, and F-measure results for each *you* type based on three class classification.

	Accuracy
Gupta + Frampton automatic	63.6%
Gupta + Frampton automatic + gesture	**66.2%**
Gupta + Frampton automatic + simple dialogue act	66.6%
Gupta + Frampton automatic + simple dialogue act + gesture	**69.0%**

Table 5: Accuracy for 3-way classification by combining gesture information with automatically extracted features based on the Decision Tree model

tioned, this level of granularity may not be practically extractable for use in a current state-of-the-art system. As a result, we implemented the simpler dialogue act incorporation method proposed by (Gupta et al., 2007a), in which only the presence or absence of the elicit dialogue act type is considered. Using this feature with the automatically extracted features yielded accuracy of 66.6%, a statistically significant improvement ($p < 0.01$) of an absolute 3% over a fully automatic run. Furthermore, if we incorporate gesture features with this configuration, the performance increases to 69.0% (statistically significantly, $p < 0.01$). This suggests that while gesture features may be redundant with information provided by the full set of dialogue act tags, it is largely complementary with the simpler dialogue act incorporation. The incorporation of gesture along with simpler and more reliable dialogue acts can potentially approach the performance gained by incorporation of more complex dialogue acts, which are often difficult to obtain. Of course, gesture features themselves are often difficult to obtain. However, redundancy in two potentially error-prone feature sources can be an asset, as data from one source may help to compensate for errors in the other. Although addressing a different problem of multimodal integration, previous work (Oviatt et al., 1997) appears to indicate that this is the case.

6 Conclusion

In this paper, we investigate the role of hand gestures in disambiguating types of *You* expressions in multiparty meetings for the purpose of machine translation.

Our results have shown that on the binary generic vs. referential classification problem, the inclusion of gesture data provides a statistically significant increase in performance over the same feature set without gesture. This result is consistent with our hypothesis that gesture data would be helpful because speakers are more likely to gesture when producing referential instances of *you*.

To produce results more akin to those that would be expected during incorporation in a real machine translation system, we experimented with the type identification problem as a three class classification problem. It was discovered that when a full set of dialogue act tags were used as features, the incorporation of gesture features does not provide an increase in performance. However, when simpler dialogue act tags are used, the incorporation of gestures helps to make up for lost performance. Since it remains a difficult problem to automatically predict complex dialog acts with high accuracy, the incorporation of gesture features may prove beneficial to current systems.

7 Acknowledgement

References

Tyler Baldwin, Joyce Y. Chai, and Katrin Kirchhoff. 2009. Communicative gestures in coreference identification in multiparty meetings. In *ICMI-MLMI '09: Proceedings of the 2009 international conference on Multimodal interfaces*, pages 211–218. ACM.

J. B. Bavelas, N. Chovil, L. Coates, and L. Roe. 1995. Gestures specialized for dialogue. *Personality and Social Psychology Bulletin*, 21:394–405.

Jacob Eisenstein and Randall Davis. 2006. Gesture improves coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 37–40, New York City, USA, June. Association for Computational Linguistics.

Jacob Eisenstein and Randall Davis. 2007. Conditional modality fusion for coreference resolution. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 352–359, Prague, Czech Republic, June. Association for Computational Linguistics.

Matthew Frampton, Raquel Fernández, Patrick Ehlen, Mario Christoudias, Trevor Darrell, and Stanley Peters. 2009. Who is "you"?: combining linguistic and gaze features to resolve second-person references in dialogue. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 273–281, Morristown, NJ, USA. Association for Computational Linguistics.

Surabhi Gupta, John Niekrasz, Matthew Purver, and Dan Jurafsky. 2007a. Resolving you in multi-party dialog. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*.

Surabhi Gupta, Matthew Purver, and Dan Jurafsky. 2007b. Disambiguating between generic and referential you in dialog. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Adam Kendon. 1980. Gesticulation and speech: Two aspects of the process of utterance. In Mary Richie Key, editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227.

D. McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

W. M. Meyers. 1990. Current generic pronoun usage. *American Speech*, 65(3):228–237.

Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 415–422, New York, NY, USA. ACM.

Andrei Popescu-Belis and Paula Estrella. 2007. Generating usable formats for metadata and annotations in a large meeting corpus. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 93–96, Prague, Czech Republic, June. Association for Computational Linguistics.

Matthew Purver, Raquel Fernández, Matthew Frampton, and Stanley Peters. 2009. Cascaded lexicalised classifiers for second-person reference resolution. In *SIGDIAL '09: Proceedings of the SIGDIAL 2009 Conference*, pages 306–309, Morristown, NJ, USA. Association for Computational Linguistics.