Incremental Reference Resolution: The Task, Metrics for Evaluation, and a Bayesian Filtering Model that is Sensitive to Disfluencies

David Schlangen, Timo Baumann, Michaela Atterer
Department of Linguistics
University of Potsdam, Germany
{das|timo|atterer}@ling.uni-potsdam.de

Abstract

In this paper we do two things: a) we discuss in general terms the task of incremental reference resolution (IRR), in particular resolution of exophoric reference, and specify metrics for measuring the performance of dialogue system components tackling this task, and b) we present a simple Bayesian filtering model of IRR that performs reasonably well just using words directly (no structure information and no hand-coded semantics): it picks the right referent out of 12 for around 50 % of real-world dialogue utterances in our test corpus. It is also able to learn to interpret not only words but also hesitations, just as humans have shown to do in similar situations, namely as markers of references to hard-to-describe entities.

1 Introduction

Like other tasks involved in language comprehension, reference resolution—that is, the linking of natural language expressions to contextually given entities—is performed incrementally by human listeners. This was shown for example by Tanenhaus et al. (1995) in a famous experiment where addressees of utterances containing referring expressions made eye movements towards target objects very shortly after the end of the first word that unambiguously specified the referent, even if that wasn't the final word of the phrase. In fact, as has been shown in later experiments (Brennan and Schober, 2001; Bailey and Ferreira, 2007; Arnold et al., 2007), such disambiguating material doesn't even have to be lexical: under certain circumstances, a speaker's hesitating already seems to be understood as increasing the likelihood of subsequent reference to hard-to-describe entities.

Recently, efforts have begun to build dialogue systems that make use of incremental processing as well (Aist et al., 2006; Skantze and Schlangen, 2009). These efforts have so far focused on aspects other than resolution of references ((Stoness et al., 2004) deals with the interaction of reference and parsing). In this paper, we discuss in general terms the task of incremental reference resolution (IRR) and specify metrics for evaluating incremental components for this task. To make the discussion more concrete, we also describe a simple Bayesian filtering model of IRR in a domain with a small number of possible referents, and show that it performs better wrt. our metrics if given information about hesitations—thus providing computational support for the rationality of including observables other than words into models of dialogue meaning.

The remainder of the paper is structured as follows: We discuss the IRR task in Section 2, and suitable evaluation metrics in Section 3. In Section 4 we describe and analyse the data for which we present results with our Bayesian model for IRR in Section 5.

2 Incremental Reference Resolution

To a first approximation, IRR can be modeled as the 'inverse' as it were of the task of generating referring expressions (GRE; which is well-studied in computational linguistics, see e. g. (Dale and Reiter, 1995)). Where in GRE words are *added* that express features which reduce the size of the set of possible distractors (with which the object that the expression is intended to pick out can be confused), in IRR words are *encountered* that express features that reduce the size of the set of possible

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 30–37,
Queen Mary University of London, September 2009.

referents. To give a concrete example, for the expression in (1-a), we could imagine that the logical representation in (1-b) is built on a word-by-word basis, and at each step the expression is checked against the world model to see whether the reference has become unique.

(1) a. the red cross
b. $\iota x(red(x) \wedge cross(x))$

To give an example, in a situation where there are available for reference only one red cross, one green circle, and two blue squares, we can say that after "the red" the referent should have been found; in a world with two red crosses, we would need to wait for further restricting information (e. g. ". . . on the left").

This is one way to describe the task, then: a component for incremental reference resolution takes expressions as input in a word-by-word fashion and delivers for each new input a set (possibly a singleton set) as output which collects those discourse entities that are compatible with the expression up to that point. (This description is meant to be neutral as to whether reference is exophoric, i. e. directly to entities in the world, or anaphoric, via previous mentions; we will mainly discuss the former case, though.)

As we will see below, this does however not translate directly into a usable metric for evaluation. While it is easy to identify the contributions of individual words in simple, constructed expressions like (1-a), reference in real conversations is often much more complex, and is a collaborative process that isn't confined to single expressions (Clark and Schaefer, 1987): referring is a pragmatic action that is not reducible to denotation. In our corpus (see below), we often find descriptions as in (2), where the speaker continuously adds (rather vague) material, typically until the addressee signals that she identified the item, or proposes a different way to describe it.

(2) Also das S Teil sieht so aus dass es ein einzelnes . Teilchen hat . dann . vier am Stück im rechten Winkel .. dazu nee . nee warte .. dann noch ein einzelnes das guckt auf der anderen Seite raus.
well, the S piece looks so that it has a single . piece . and then . four together in a 90 degree angle .. and also . no .. wait .. and then a single piece that sticks out on the other side.

While it's difficult to say in the individual case what the appropriate moment is to settle on a hypothesis about the intended referent, and what the "correct" time-course of the development of hypotheses is, it's easy to say what we want to be true in general: we want a referent to be found as early as possible, with as little change of opinion as possible during the utterance.[1] Hence a model that finds the correct referent earlier and makes fewer wrong decisions than a competing one will be considered better. The metrics we develop in the next section spell out this idea.

3 Evaluation Metrics for IRR

In previous work, we have discussed metrics for evaluating the performance of incremental speech recognition (Baumann et al., 2009). There, our metrics could rely on time-aligned gold-standard information against which the incremental results could be measured. For the reasons discussed in the previous section, we do not assume that we have such temporally-aligned information for evaluating IRR. Our measures described here simply assume that there is one intention behind the referring utterances (namely to identify a certain entity), and that this intention is there from the beginning of the utterance and stays constant.[2] This is not to be understood as the claim that it is reasonable to expect an IRR component to pick out a referent even if the only part of the utterance that has already been processed for example is "now take the"—it just facilitates the "earlier is better" ranking discussed above.

We use two kinds of metrics for IRR: *positional metrics*, which measure when (which percentage into the utterance) a certain event happens, and *edit metrics* which capture the "jumpiness" of the decision process (how often the component changes its mind during an utterance).

Figure 1 shows a constructed example that il-

[1] We leave open here what "as early as possible" means—a well-trained model might be able to resolve a reference before the speaker even deems that possible, and hence appear to do unnatural (or supernatural?) 'mind reading'. Conversely, frequent changes of opinion might be something that human listeners would exhibit as well (e. g. in their gaze direction). We abstract away from these finer details in our heuristic.

[2] Note that our metrics would also work for corpora where the correct point-of-identification is annotated; this would simply move the reference point from the beginning of the utterance to that point. Gallo et al. (2007) describe an annotation effort in a simpler domain where entities can easily be described which would make such information available.

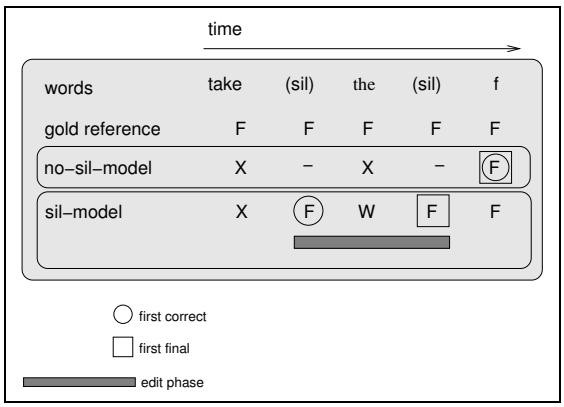

Figure 1: Simple constructed example that illustrates the evaluation measures

lustrates these ideas. We assume that reference is to an object that is internally represented by the letter F. The example shows two models, no-sil and sil (what exactly they are doesn't matter for now). The former model guesses that reference is to object X already after the first word, and stays with this opinion until it encounters the final word, when it chooses F as most likely referent. (Why the decision for the items *sil* is "-" will be explained below; here this can be read as "repetition of previous decision".) The other model changes its mind more often, but also is correct for the first time earlier and stays correct earlier. Our metrics make this observation more precise:

• **average fc** (first correct): how deep into the utterance do we make the first correct guess? (If the decision component delivers n-best lists instead of single guesses, "correct" means here and below "is member of n-best list".)

E. g., if the referent is recognised only after the final word of the expression, the score for this metric would be 1. In our example it is 2/5 for the sil-model and 1 for the non-sil model.

• **fc applicable**: since the previous measure can only be specified for cases where the correct referent has been found, we also specify for how many utterances this is the case.

• **average ff** (first final): how deep into the utterance do we make the correct guess *and* don't subsequently change our mind? This would be 4/5 for the sil-model in our example and 1 for the no-sil-model.

• **ff applicable**: again, the previous measure can only be given where the final guess of the component is correct, so we also need to specify how often this is the case. Note that whenever ff is applicable, fc is applicable as well, so **ff applicable**$\leq$**fc applicable**.

• **ed-utt** (mean edits per utterance): an IRR module may still change its mind even after it has already made a correct guess. This metric measures how often the module changes its mind before it comes back to the right guess (if at all). Since such decision-revisions (edits) may be costly for later modules, which possibly need to retract their own hypotheses that they've built based on the output of this module, ideally this number should be low.

In our example the number of edits between fc and ff is 2 for the sil-model and 0 for the non-sil model (because here fc and ff are at the same position).

• **eo** (edit overhead): ratio unnecessary edits / necessary edits. (In the ideal case, there is exactly one edit, from "no decision" to the correct guess.)

• **correctness**: how often the model guesses correctly. This is 3/5 for the sil-model in the example and 1/5 for the non-sil-model.

• **sil-correctness**: how often the model guesses correctly *during hesitations*. The correctness measure applied only to certain data-points; we use this to investigate whether informing the model about hesitations is helpful.

• **adjusted error**: some of our IRR models can return "undecided" as reply. The correctness measures defined above would punish this in the same way as a wrong guess. The **adjusted error** measure implements the idea that undecidedness is better than a wrong guess, at least early in the utterance. More precisely, it's defined to be 0 if the guess is correct, $pos \,/\, pos_{max}$ if the reply is "undecided" (with pos denoting the position in the utterance), and 1 if the guess is incorrect. That way uncertainty is not punished in the beginning of the utterance and counted like an error towards its end.

Note that these metrics characterise different aspects of the performance of a model. In practical cases, they may not be independent from each other, and a system designer will have to decide which one to optimize. If it is helpful to be informed about a likely referent early, for example to prepare a reaction, and is not terribly costly to later have to revise hypotheses, then a low **first correct** may be the target. If hypothesis revisions are costly, then a low **edit overhead** may be preferred over a low **first correct**. (**first final** and **ff applicable**, however, are parameters that are useful for global optimisation.)

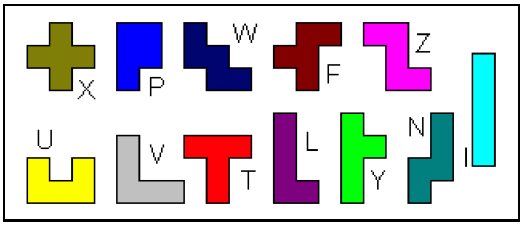

Figure 2: The Twelve Pentomino Pieces with their canonical names (which were not known to the dialogue participants). The pieces used in the dialogues all had the same colour.

In the remaining sections, we describe a probabilistic model of IRR that we have implemented, and evaluate it in terms of these metrics. We begin with describing the data from which we learnt our model.

4 Data

4.1 Our Corpora

As the basis for training and testing of our model we used data from three corpora of task-oriented dialogue that differ in some details of the set-up, but use the same task: an Instruction Giver (IG) instructs an Instruction Follower (IF) on which puzzle pieces (from the "Pentomino" game, see Figure 2) to pick up. In detail, the corpora were:

• The Pento Naming corpus described in (Siebert and Schlangen, 2008). In this variant of the task, IG *records* instructions for an absent IF; so these aren't fully interactive dialogues. The corpus contained 270 utterances out of which we selected those 143 that contained descriptions of puzzle pieces (and not of their position on the gameboard).

• Selections from the FTT/PTT corpus described in (Fernández et al., 2007), where IF and IG are connected through an audio-only connection, and in some dialogues a simplex / push-to-talk one. We selected all utterances from IG that contained references to puzzle pieces (286 altogether).

• The third part of our corpus was constructed specifically for the experiments described here. We set-up a Wizard of Oz experiment where users were given the task to describe puzzle pieces for the "dialogue system" to pick up. The system (i. e. the wizard) had available a limited number of utterances and hence could conduct only a limited form of dialogue. We collected 255 utterances containing descriptions of puzzle pieces in this way.

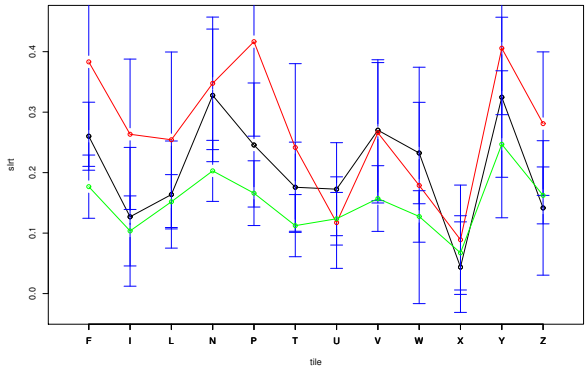

Figure 3: Silence rate per referent and corpus (WOz:black, PentoNaming:red, FTT:green)

All utterances were hand-transcribed and the transcriptions were automatically aligned with the speech data using the MAUS system (Schiel, 2004); this way, we could automatically identify pauses during utterances and measure their length. For some experiments (see below), pauses were "re-ified" through the addition of silence pseudo-words (one for each 333 ms of silence).

The resulting corpus is not fully balanced in terms of available material for the various pieces or contributions by sub-corpora.

4.2 Descriptive Statistics

We were interested to see whether intra-utterance silences (hesitations) could potentially be used as an information source in our (more or less) real-world data in the same way as was shown in the much more controlled situations described in the psycholinguistics literature mentioned above in the introduction (Arnold et al., 2007). Figure 3 shows the mean ratio of within-utterance silences per word for the different corpora and different referents. We can see that there are clear differences between the pieces. For example, references to the piece whose canonical name is X contain very few or short hesitations, whereas references to Y tend to contain many. We can also see that the tendencies seem to be remarkably similar between corpora, but with relatively stable offsets between them, PentoDescr having the longest, PTT/FTT the shortest silences. We speculate that this is the result of the differing degrees of interactivity (none in PentoDescr, restricted in WOz, less restricted in PTT, free in FTT) which puts different pressures on speakers to avoid silences. To balance our data with respect to this difference, we performed some experiments with adjusted data

where silence lengths in PentoDescr were adjusted by 0.7 and in PTT/FTT by 1.3. This brings the silence rates in the corpora, if plotted in the style of Figure 3, almost in congruence.

To test whether the differences in silence rate between utterances referring to different pieces are significant, we performed an ANOVA and found a main effect of silence rate, $F(11, 672) = 6.2102, p < 8.714^{-10}$. A post-hoc t-test reveals that there are roughly two groups whose members are not significantly different within-group, but are across groups: I, L, U, W and X form one group with relatively low silence rate, F, N, P, T, V, Y, and Z another with relatively high silence rate. We will see in the next section whether our model picked up on these differences.

5 A Bayesian Filtering Model of IRR

To explore incremental reference resolution, and as part of a larger incremental dialogue system we are building, we implemented a probabilistic reference resolver that works in the pentomino domain. At its base, the resolver has a Bayesian Filtering model (see e. g. (Thrun et al., 2005)) that with each new observation (word) computes a belief distribution over the available objects (the twelve puzzle pieces); in a second step, a decision for a piece (or a collection of pieces in the n-best case) is derived from this distribution. This model is incremental in a very natural and direct way: new input increments are simply treated as new observations that update the current belief state. Note that this model does not start with any assumptions about semantic word classes: whether an observed word carries information about what is being referred to will be learnt from data.

5.1 The Belief-Update Model

We use a Bayesian model which treats the intended referent as a latent variable generating a sequence of observations ($w_{1:n}$ is the sequence of words $w_1, w_2, \ldots, w_n$):
$P(r|w_{1:n}) = \alpha * P(w_n|r, w_{1:n-1}) * P(r|w_{1:n-1})$
where
• $P(w_n|r, w_{1:n-1})$ is the likelihood of the new observation (see below for how we approximate that); and
• the prior $P(r|w_{1:n-1})$ at step n is the posterior of the previous step. Before the first observation is made (i. e., the first word is seen), the prior is simply a distribution over the possible referents, $P(r)$.

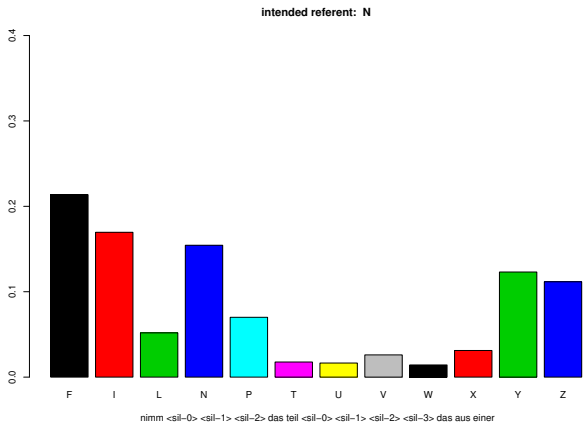

Figure 4: Example of Belief Distribution after Observation

In our experiment, we set this to a uniform distribution, but if there is prior information from other sources (e. g., because the dialogue state makes certain pieces more salient), this can be reflected.
• α is a normalising constant, ensuring that the result is indeed a probability distribution.

The output of the model is a distribution of belief over the 12 available entities, as shown in Figure 4. Figure 5 shows in a 3D plot the development of the belief state (pieces from front to back, strength of belief as height of the peaks) over the course of a whole utterance (with observations from left to right).

5.2 The Decision Step

We implemented several ways to derive a decision for a referent from such a distribution:

i) In the *arg max* approach, at each state the referent with the highest posterior probability is chosen. For Figure 4, that would be F (and hence, a wrong decision). As Figure 5 shows (and the example is quite representative for the model behaviour), there often are various local maxima over the course of an utterance, and hence a model that takes as its decision always the maximum can be expected to perform many edits.

ii) In the *adaptive threshold* approach, we start with a default decision for a special 13th class, "undecided", and a new decision is only made if the maximal value at the current step is above a certain threshold, where this threshold is reset every time this condition is met. In other words, this draws a plane into the belief space and only makes a new decision when a peak rises above this plane and hence above the previous peak. In effect, this approach favours strong convictions and reduces

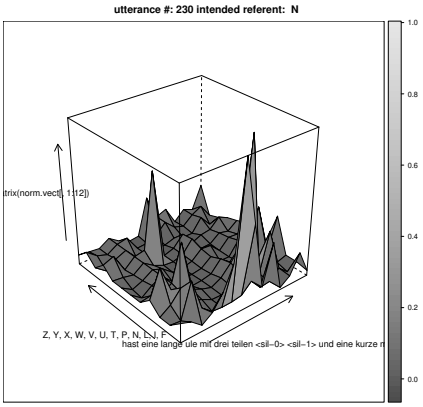

Figure 5: Belief Update over Course of Utterance

the "jitter" in the decision making.

In our example from Figure 4, this would mean that the maximum, F, would only be the decision if its value was higher than the threshold and there was no previous guess that was even higher.

iii) The final model implements a *threshold n-best* approach, where not just a single piece is selected but all pieces that are above a certain threshold. Assuming that the threshold is 0.1 for example this would select F, I, N, Y, and Z—and hence would include the correct reference in Figure 4.

5.3 Implementation

To learn and query the observation likelihoods $P(w_n|r, w_{1:n-1})$, we used referent-specific language models. More precisely, we computed the likelihood as $P(r, w_{1:n})/P(r, w_{1:n-1})$ (definition conditional probability), and approximated the joint probabilities of referent and word sequence via n-grams with specialised words. E. g., an utterance like "take the long, narrow piece" referring to piece I (or tested for reference to this piece) would be rewritten as "take_I the_I long_I narrow_I piece_I" and presented to the n-gram learner / inference component. (Both taken from the SRI LM package, (Stolcke, 2002).)

During evaluation of the models, the test utterances are fed word-by-word to the model and the decision is evaluated against the known intended referent. Since we were interested in testing whether disfluencies contained information that would be learned, for one variant of the system we also fed pseudo-words for silences and hesitation markers like *uhm*, numbered by their position (i. e., "take the .." becomes "take the sil-1 sil-2"), to both learning and inference for the *silence-sensitive* variant; the *silence-ignorant* variant simply repeats the previous decision at such points and does not update its belief state; this way, it is guaranteed that both variants generate the same number of decisions and can be compared directly. (Cf. the dashes in the "no-sil-model" in Figure 1 above: those are points where no real computation is made in the no-sil case.)

5.4 Experiments

All experiments were performed with 10-fold cross-validation. We always ran both versions, the one that showed silences to the model and the one that didn't. We tested various combinations of language model parameters and deciders, of which the best-performing ones are discussed in the next section.

5.5 Results

Table 1 shows the results for the different decision methods and for models where silences are included as observations and where they aren't, and, as a baseline, the result for a resolver that makes a random decision after each observation.

As we can see, the different decision methods have different characteristics wrt. individual measures. The *threshold n-best* approach performs best across the board—but of course has a slightly easier job since it does not need to make unambiguous decisions. We will look into the development of the n-best lists in a second, but for now we note that this model is for almost all utterances correct at least once (97 % **fc applicable**) and if so, typically very early (after 30 % of the utterance). In over half of the cases (54.68 %), the final decision is correct (i. e. is an n-best list that contains the correct referent), and similarly for a good third of all silence observations. Interestingly, silence-correctness is decidedly higher for the silence model (which does actually make new decisions during silences and hence based on the information that the speaker is hesitating) than for the non-sil model (which at these places only repeats the previously made decision). The model performs significantly bettern than a baseline that randomly selects n-best lists of the same size (see *rnd-nb* in Table 1).

As can be expected, the *adaptive threshold* approach is more stable with its decisions, as witnessed by the low **edit overhead**. The fact that it changes its decision not as often has an impact on the other measures, though: in more cases, the model is correct not even once (**fc applicable** is

	n-best		rnd-nb	adapt		max		random
Measure / Model	w/ h	w/o h	w/ h	w/ h	w/o h	w/ h	w/o h	w/ h
fc applicable	97.22 %	95.03 %	85.38 %	63.15 %	66.67 %	86.55 %	82.89 %	59.94 %
average fc	30.43 %	33.73 %	29.61 %	53.87 %	55.25 %	46.55 %	49.31 %	42.60 %
ff applicable	54.68 %	54.24 %	17.54 %	48.68 %	53.07 %	39.77 %	40.64 %	9.65 %
average ff	87.74 %	85.01 %	97.08 %	71.24 %	70.89 %	96.08 %	94.28 %	98.44 %
edit overhead	93.49 %	90.65 %	96.65 %	69.61 %	67.66 %	92.57 %	89.44 %	93.16 %
correctness	37.81 %	36.81 %	23.37 %	23.01 %	26.61 %	17.83 %	20.23 %	7.83 %
sil-correctness	36.60 %	31.09 %	26.39 %	18.71 %	22.58 %	13.67 %	19.34 %	8.63 %
adjusted error	60.07 %	56.96 %	76.63 %	76.29 %	70.90 %	82.17 %	79.42 %	92.16 %

Table 1: Results for different decision methods (*n-best*, *adapt*ive, *max* arg and *random*) and for models with and without silence-observations (*w/ h* and *w/o h*, respectively)

lower than for the other two models). But it is still correct with almost half of its final decisions, and these come even earlier than for the n-best model. Silence information does not seem to help this model; this suggests that the information provided by knowledge about the fact that the speaker hesitates is too subtle to push through the threshold in order to change decisions.

The *arg max* approach fares worst. Since neither the relative strength of the strongest belief (as compared to that in the competing pieces) nor the global strength (have I been more convinced before?) is taken into account, the model changes its mind too often, as evidenced by the edit overhead, and does not settle on the correct referent often (and if, then late). Again, silence information does not seem to be helpful for this model.

As a more detailed look at what happens during silence sequences, Figure 6 plots the average change in probability from onset of silence to a point at 1333 ms of silence. (Recall that the underlying Bayesian model is the same for all models evaluated above, they differ only in how they derive a decision.) We can see that the gains and losses are roughly as expected from the analysis of the corpora: pieces like L and P become more expected after a silence of that length, pieces like X less. So the model does indeed seem to learn that hesitations systematically occur together with certain pieces. (The reader can convince herself with the help of Figure 2 that these shapes are indeed comparatively hard-to-describe; but the interesting point here is that this categorisation does not have to be brought to the model but rather is discovered by it.)

Finally, a look at the distribution and the sizes of the n-best groupings: the most frequent decision is

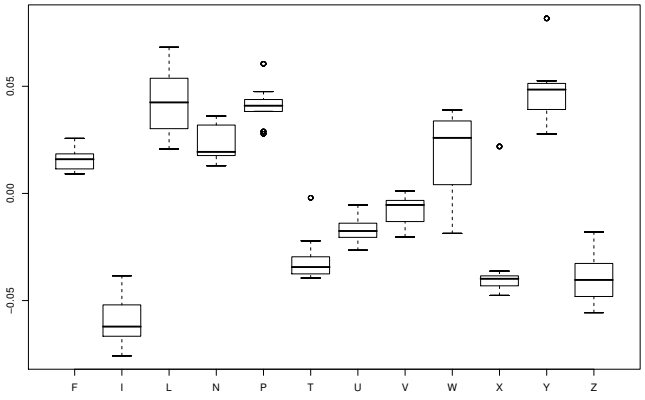

Figure 6: Average change in probability from onset of silence to 1333 ms into silence

"undecided" (474 times), followed by the groupings F_N, N_Y, and N_Y_P (343, 342 and 196, respectively). Here again we find groupings that reflect the differences w.r.t. hesitation rate. The average size of the n-best lists is 2.58 (sd = 1.4).

6 Conclusions and Further Work

We discussed the task of incremental reference resolution (IRR), in particular with respect to exophoric reference. From a theoretical perspective, it might seem easy to specify what the ideal behaviour of an IRR component should be, namely to always produce the set of entities (the extension) that is compatible with the part of the expression seen so far. In practice, however, this is difficult to annotate, for both practical reasons as well as theoretical (referring is a pragmatic activity that is not reducible to denotation). The metrics we defined for evaluation of IRR components account for this in that they do not require a gold

standard annotation that fixes the dynamics of the resolution process; they simply make it possible to quantify the assumption that "early and with strong convictions" is best.

We then presented our probabilistic model of IRR that works directly on word observations without any further processing (POS tagging, parsing). It achieves a reasonable success (as measured with our metrics); for example, in over half of the cases, the final guess of the model is correct, and comes before the utterance is over. As an additional interesting feature, the model is able to interpret hesitations (silences lifted to pseudo-word status) in a way shown before only in controlled psycholinguistic experiments, namely as making reference to hard-to-describe pieces more likely.[3]

In future work, we want to explore the model's performance on ASR output. It is not clear a priori that this would degrade performance much, as it can be expected that the learning components are quite robust against noise. Connected to this, we want to explore more complex statistical models, e. g. a hierarchical model where one level generates parts of the utterance (e. g. non-referential parts and referential parts) and the second the actual words. We also want to test how this approach scales up to worlds with a larger number of possible referents, where consequently approximation methods like particle filtering have to be used. Finally, we will test how the module contributes to a working dialogue system, where further decisions (e. g. for clarification requests) can be built on its output.

Acknowledgments This work was funded by a grant from DFG in the Emmy Noether Programme. We would like to thank the anonymous reviewers for their detailed comments.

References

G.S. Aist, J. Allen, E. Campana, L. Galescu, C.A. Gomez Gallo, S. Stoness, M. Swift, and M Tanenhaus. 2006. Software architectures for incremental understanding of human speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, PA, USA, September.

Jennifer E. Arnold, Carla L. Hudson Kam, and Michael K. Tanenhaus. 2007. If you say *thee uh* you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology*.

Karl Bailey and F. Ferreira. 2007. The processing of filled pause disfluencies in the visual world. In R. P. G. von Gompel, M H. Fischer, W. S. Murray, and R. L. Hill, editors, *Eye Movements: A Window on Mind and Brain*, chapter 22. Elsevier.

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of NAACL-HLT 2009*, Boulder, USA.

Susan E. Brennan and Michael F. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44:274–296.

Herbert H. Clark and Edward F. Schaefer. 1987. Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2(1):19–41.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19:233–263.

Raquel Fernández, David Schlangen, and Tatjana Lucht. 2007. Push-to-talk ain't always bad! comparing different interactivity settings in task-oriented dialogue. In *Proceeding of DECALOG (SemDial'07)*, Trento, Italy, June.

Carlos Gómez Gallo, Gregory Aist, James Allen, William de Beaumont, Sergio Coria, Whitney Gegg-Harrison, Joana Paulo Pardal, and Mary Swift. 2007. Annotating continuous understanding in a multimodal dialogue corpus. In *Proceeding of DECALOG (SemDial07)*, Trento, Italy, June.

Florian Schiel. 2004. Maus goes iterative. In *Proc. of the IV. International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Alexander Siebert and David Schlangen. 2008. A simple method for resolution of definite reference in a shared visual context. In *Procs of SIGdial*, Columbus, Ohio.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of EACL 2009*, Athens, Greece, April.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings Intl. Conf. Spoken Language Processing (ICSLP'02)*, Denver, Colorado, USA, September.

Scott C. Stoness, Joel Tetreault, and James Allen. 2004. Incremental parsing with reference interaction. In *Proceedings of the Workshop on Incremental Parsing at the ACL 2004*, pages 18–25, Barcelona, Spain, July.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathllen M. Eberhard, and Julie C. Sedivy. 1995. Intergration of visual and linguistic information in spoken language comprehension. *Science*, 268.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. 2005. *Probabilistic Robotics*. MIT Press, Cambridge, Massachusetts, USA.

[3] It is interesting to speculate whether this could have implications for generation of referring expressions as well. It might be a good strategy to make your planning problems observable or even to fake planning problems that are understandable to humans.