Towards Finding and Fixing Fragments: Using ML to Identify Non-Sentential Utterances and their Antecedents in Multi-Party Dialogue

David Schlangen
Department of Linguistics
University of Potsdam
P.O. Box 601553
D-14415 Potsdam — Germany
das@ling.uni-potsdam.de

Abstract

Non-sentential utterances (e.g., short-answers as in "Who came to the party?"—"Peter.") are pervasive in dialogue. As with other forms of ellipsis, the elided material is typically present in the context (e.g., the question that a short answer answers). We present a machine learning approach to the novel task of identifying fragments and their antecedents in multi-party dialogue. We compare the performance of several learning algorithms, using a mixture of structural and lexical features, and show that the task of identifying antecedents given a fragment can be learnt successfully ($f(0.5) = .76$); we discuss why the task of identifying fragments is harder ($f(0.5) = .41$) and finally report on a combined task ($f(0.5) = .38$).

1 Introduction

Non-sentential utterances (NSUs) as in (1) are pervasive in dialogue: recent studies put the proportion of such utterances at around 10% across different types of dialogue (Fernández and Ginzburg, 2002; Schlangen and Lascarides, 2003).

(1) a. A: Who came to the party?
B: Peter. (= *Peter came to the party.*)

b. A: I talked to Peter.
B: Peter Miller? (= *Was it Peter Miller you talked to?*)

c. A: Who was this? Peter Miller? (= *Was this Peter Miller?*

Such utterances pose an obvious problem for natural language processing applications, namely that the intended information (in (1-a)-B a proposition) has to be recovered from the uttered information (here, an NP meaning) with the help of information from the context.

While some systems that automatically resolve such fragments have recently been developed (Schlangen and Lascarides, 2002; Fernández et al., 2004a), they have the drawback that they require "deep" linguistic processing (full parses, and also information about discourse structure) and hence are not very robust. We have defined a well-defined subtask of this problem, namely identifying *fragments* (certain kinds of NSUs, see below) and their antecedents (in multi-party dialogue, in our case), and present a novel machine learning approach to it, which we hypothesise will be useful for tasks such as automatic meeting summarisation.[1]

The remainder of this paper is structured as follows. In the next section we further specify the task and different possible approaches to it. We then describe the corpus we used, some of its characteristics with respect to fragments, and the features we extracted from it for machine learning. Section 4 describes our experimental settings and reports the results. After a comparison to related work in Section 5, we close with a conclusion and some further

[1] (Zechner and Lavie, 2001) describe a related task, linking questions and answers, and evaluate its usefulness in the context of automatic summarisation; see Section 5.

Proceedings of the 43rd Annual Meeting of the ACL, pages 247–254,
Ann Arbor, June 2005.

work that is planned.

2 The Tasks

As we said in the introduction, the main task we want to tackle is to align (certain kinds of) NSUs and their *antecedents*. Now, what characterises this kind of NSU, and what are their antecedents?

In the examples from the introduction, the NSUs can be resolved simply by looking at the previous utterance, which provides the material that is elided in them. In reality, however, the situation is not that simple, for three reasons: First, it is of course not always the *previous* utterance that provides this material (as illustrated by (2), where utterance 7 is resolved by utterance 1); in our data the average distance in fact is 2.5 utterances (see below).

(2)
	1	B:	[. . .] What else should be done ?
	2	C:	More intelligence .
	3		More good intelligence .
	4		Right .
	5	D:	Intelligent intelligence .
	6	B:	Better application of face and voice recognition .
	7	C:	More [. . .] intermingling of the agencies , you know .

[from NSI 20011115]

Second, it's not even necessarily a single utterance that does this–it might very well be a span of utterances, or something that has to be inferred from such spans (parallel to the situation with pronouns, as discussed empirically e.g. in (Strube and Müller, 2003)). (3) shows an example where a new topic is broached by using an NSU. It is possible to analyse this as an answer to the *question under discussion* "what shall we organise for the party?", as (Fernández et al., 2004a) would do; a question, however, which is only *implicitly* posed by the previous discourse, and hence this is an example of an NSU that does not have an overt antecedent.

(3) [after discussing a number of different topics]
1	D:	So, equipment.
2		I can bring [. . .]

[from NSI 20011211]

Lastly, not all NSUs should be analysed as being the result of ellipsis: backchannels for example (like the "Right" in utterance 4 in (2) above) seem to directly fulfil their discourse function without any need for reconstruction.[2]

To keep matters simple, we concentrate in this paper on NSUs of a certain kind, namely those that a) do not predominantly have a discourse-management function (like for example backchannels), but rather convey messages (i.e., propositions, questions or requests)—this is what distinguishes *fragments* from other NSUs—and b) have individual utterances as antecedents. In the terminology of (Schlangen and Lascarides, 2003), fragments of the latter type are *resolution-via-identity*-fragments, where the elided information can be identified in the context and need not be inferred (as opposed to *resolution-via-inference*-fragments). Choosing only this special kind of NSUs poses the question whether this subgroup is distinguished from the general group of fragments by criteria that can be learnt; we will return to this below when we analyse the errors made by the classifier.

We have defined two approaches to this task. One is to split the task into two sub-tasks: identifying fragments in a corpus, and identifying antecedents for fragments. These steps are naturally performed sequentially to handle our main task, but they also allow the fragment classification decision to come from another source—a language-model used in an automatic speech recognition system, for example—and to use only the antecedent-classifier. The other approach is to do both at the same time, i.e. to classify pairs of utterances into those that combine a fragment and its antecedent and those that don't. We report the results of our experiments with these tasks below, after describing the data we used.

3 Corpus, Features, and Data Creation

3.1 Corpus

As material we have used six transcripts from the "NIST Meeting Room Pilot Corpus" (Garofolo et al., 2004), a corpus of recordings and transcriptions of multi-party meetings.[3] Those six transcripts con-

[2] The boundaries are fuzzy here, however, as backchannels can also be fragmental repetitions of previous material, and sometimes it is not clear how to classify a given utterance. A similar problem of classifying fragments is discussed in (Schlangen, 2003) and we will not go further into this here.

[3] We have chosen a multi-party setting because we are ultimately interested in automatic summarisation of meetings. In this paper here, however, we view our task as a "stand-alone task". Some of the problems resulting in the presence of many

average distance $\alpha - \beta$ (utterances):	2.5
α declarative	159 (52%)
α interrogative	140 (46%)
α unclassfd.	8 (2%)
β declarative	235 (76%)
β interrogative	(23%)
β unclassfd.	2 (0.7%)
α being last in their turn	142 (46%)
β being first in their turn	159 (52%)

Table 1: Some distributional characteristics. (α denotes antecedent, β fragment.)

sist of 5,999 utterances, among which we identified 307 fragment–antecedent pairs.[4,5] With 5.1% this is a lower rate than that reported for NSUs in other corpora (see above); but note that as explained above, we are actually only looking at a sub-class of all NSUs here.

For these pairs we also annotated some more attributes, which are summarised in Table 1. Note that the average distance is slightly higher than that reported in (Schlangen and Lascarides, 2003) for (2-party) dialogue (1.8); this is presumably due to the presence of more speakers who are able to reply to an utterance. Finally, we automatically annotated all utterances with part-of-speech tags, using `TreeTagger` (Schmid, 1994), which we've trained on the switchboard corpus of spoken language (Godfrey et al., 1992), because it contains, just like our corpus, speech disfluencies.[6]

We now describe the creation of the data we used for training. We first describe the data-sets for the different tasks, and then the features used to represent the events that are to be classified.

3.2 Data Sets

Data creation for the fragment-identification task (henceforth simply *fragment-task*) was straightforward: for each utterance, a number of features was derived automatically (see next section) and the correct class (fragment / other) was added. (Note that none of the manually annotated attributes were used.) This resulted in a file with 5,999 data points for classification. Given that there were 307 fragments, this means that in this data-set there is a ratio positives (fragments) vs. negatives (non-fragments) for the classifier of 1:20. To address this imbalance, we also ran the experiments with balanced data-sets with a ratio of 1:5.

The other tasks, antecedent-identification (*antecedent-task*) and antecedent-fragment-identification (*combined-task*) required the creation of data-sets containing pairs. For this we created an "accessibility window" going back from each utterance. Specifically, we included for each utterance a) all previous utterances of the same speaker from the same turn; and b) the three last utterances of every speaker, but only until one speaker took the turn again and up to a maximum of 6 previous utterances. To illustrate this method, given example (2) it would form pairs with utterance 7 as fragment-candidate and all of utterances 6–2, but not 1, because that violates condition b) (it is the second turn of speaker B).

In the case of (2), this exclusion would be a wrong decision, since 1 is in fact the antecedent for 7. In general, however, this dynamic method proved good at capturing as many antecedents as possible while keeping the number of data points manageable. It captured 269 antecedent-fragment pairs, which had an average distance of 1.84 utterances. The remaining 38 pairs which it missed had an average distance of 7.27 utterances, which means that to capture those we would have had to widen the window considerably. E.g., considering all previous 8 utterances would capture an additional 25 pairs, but at the cost of doubling the number of data points. We hence chose the approach described here, being aware of the introduction of a certain bias.

As we have said, we are trying to link *utterances*, one a fragment, the other its antecedent. The notion of *utterance* is however less well-defined than one might expect, and the segmentation of continuous speech into utterances is a veritable research problem on its own (see e.g. (Traum and Heeman, 1997)). Often it is arguable whether a prepositional

speakers are discussed below.

[4] We have used the MMAX tool (Müller and Strube, 2001)) for the annotation.

[5] To test the reliability of the annotation scheme, we had a subset of the data annotated by two annotators and found a satisfactory κ-agreement (Carletta, 1996) of $\kappa = 0.81$.

[6] The tagger is available free for academic research from `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`.

Structural features	
dis	distance α – β, in utterances
sspk	same speaker yes/no
nspk	number speaker changes (= # turns)
iqu	number of intervening questions
alt	α last utterance in its turn?
bft	β first utterance in its turn?

Lexical / Utterance-based features	
bvb	(tensed) verb present in β?
bds	disfluency present in β?
aqm	α contains question mark
awh	α contains wh word
bpr	ratio of polar particles (*yes*, *no*, *maybe*, etc..) / other in β
apr	ratio of polar particles in α
lal	length of α
lbe	length of β
nra	ratio nouns / non-nouns in α
nra	ratio nouns / non-nouns in β
rab	ratio nouns in β that also occur in α
rap	ratio words in β that also occur in α
god	google similarity (see text)

Table 2: The Features

phrase for example should be analysed as an adjunct (and hence as not being an utterance on its own) or as a fragment. In our experiments, we have followed the decision made by the transcribers of the original corpus, since they had information (e.g. about pauses) which was not available to us.

For the antecedent-task, we include only pairs where β (the second utterance in the pair) is a fragment—since the task is to identify an antecedent for already identified fragments. This results in a data-set with 1318 data points (i.e., we created on average 4 pairs per fragment). This data-set is sufficiently balanced between positives and negatives, and so we did not create another version of it. The data for the combined-task, however, is much bigger, as it contains pairs for all utterances. It consists of 26,340 pairs, i.e. a ratio of roughly 1:90. For this reason we also used balanced data-sets for training, where the ratio was adjusted to 1:25.

3.3 Features

Table 2 lists the features we have used to represent the utterances. (In this table, and in this section, we denote the candidate for being a fragment with β and the candidate for being β's antecedent with α.)

We have defined a number of structural features, which give information about the (discourse-)structural relation between α and β. The rationale behind choosing them should be clear; iqu for example indicates in a weak way whether there might have been a topic change, and high nspk should presumably make an antecedent relation between α and β less likely.

We have also used some lexical or utterance-based features, which describe lexical properties of the individual utterances and lexical relations between them which could be relevant for the tasks. For example, the presence of a verb in β is presumably predictive for its being a fragment or not, as is the length. To capture a possible semantic relationship between the utterances, we defined two features. The more direct one, rab, looks at verbatim re-occurrences of nouns from α in β, which occur for example in check-questions as in (4) below.

(4) A: I saw Peter.
B: Peter? (= *Who is this Peter you saw?*)

Less direct semantic relations are intended to be captured by god, the second semantic feature we use.[7] It is computed as follows: for each pair (x, y) of nouns from α and β, Google is called (via the Google API) with a query for x, for y, and for x and y together. The similarity then is the average ratio of pair vs. individual term:

$$\textit{Google\_Similarity}(x,y) = (\frac{hits(x,y)}{hits(x)} + \frac{hits(x,y)}{hits(y)}) * \frac{1}{2}$$

We now describe the experiments we performed and their results.

4 Experiments and Results

4.1 Experimental Setup

For the learning experiments, we used three classifiers on all data-sets for the the three tasks:

- SLIPPER (Simple Learner with Iterative Pruning to Produce Error Reduction), (Cohen and Singer, 1999), which is a rule learner which combines the separate-and-conquer approach with confidence-rated boosting. It is unique among the classifiers that

[7] The name is short for *google distance*, which indicates its relatedness to the feature used by (Poesio et al., 2004); it is however a measure of *similarity*, not distance, as described above.

we have used in that it can make use of "set-valued" features, e.g. strings; we have run this learner both with only the features listed above and with the utterances (and POS-tags) as an additional feature.

• TiMBL (Tilburg Memory-Based Learner), (Daelemans et al., 2003), which implements a memory-based learning algorithm (IB1) which predicts the class of a test data point by looking at its distance to all examples from the training data, using some distance metric. In our experiments, we have used the weighted-overlap method, which assigns weights to all features.

• MaxEnt, Zhang Le's C++ implementation[8] of maximum entropy modelling (Berger et al., 1996). In our experiments, we used L-BFGS parameter estimation.

We also implemented a naïve bayes classifier and ran it on the fragment-task, with a data-set consisting only of the strings and POS-tags.

To determine the contribution of all features, we used an iterative process similar to the one described in (Kohavi and John, 1997; Strube and Müller, 2003): we start with training a model using a baseline set of features, and then add each remaining feature individually, recording the gain (w.r.t. the f-measure ($f(0.5)$, to be precise)), and choosing the best-performing feature, incrementally until no further gain is recorded. All individual training- and evaluation-steps are performed using 8-fold cross-validation (given the small number of positive instances, more folds would have made the number of instances in the test set set too small).

The baselines were as follows: for the fragment-task, we used `bvb` and `lbe` as baseline, i.e. we let the classifier know the length of the candidate and whether the candidate contains a verb or not. For the antecedent-task we tested a very simple baseline, containing only of one feature, the distance between α and β (`dis`). The baseline for the combined-task, finally, was a combination of those two baselines, i.e. `bvb+lbe+dis`. The full feature-set for the fragment-task was `lbe, bvb, bpr, nrb, bft, bds` (since for this task there was no α to compute features of), for the two other tasks it was the complete set shown in Table 2.

[8] Available from `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`.

4.2 Results

The Tables 3–5 show the results of the experiments. The entries are roughly sorted by performance of the classifier used; for most of the classifiers and data-sets for each task we show the performance for baseline, intermediate feature set(s), and full feature-set, for the rest we only show the best-performing setting. We also indicate whether a balanced or unbalanced data set was used. I.e., the first three lines in Table 3 report on MaxEnt on a balanced data set for the fragment-task, giving results for the baseline, baseline+`nrb+bft`, and the full feature-set.

We begin with discussing the fragment task. As Table 3 shows, the three main classifiers perform roughly equivalently. Re-balancing the data, as expected, boosts recall at the cost of precision. For all settings (i.e., combinations of data-sets, feature-sets and classifier), except re-balanced maxent, the baseline (verb in β yes/no, and length of β) already has some success in identifying fragments, but adding the remaining features still boosts the performance. Having available the string (condition s.s; slipper with set valued features) interestingly does not help SLIPPER much.

Overall the performance on this task is not great. Why is that? An analysis of the errors made shows two problems. Among the false negatives, there is a high number of fragments like "yeah" and "mhm", which in their particular context were answers to questions, but that however occur much more often as backchannels (true negatives). The classifier, without having information about the context, can of course not distinguish between these cases, and goes for the majority decision. Among the false positives, we find utterances that are indeed non-sentential, but for which no antecedent was marked (as in (3) above), i.e., which are not fragments in our narrow sense. It seems, thus, that the required distinctions are not ones that can be reliably learnt from looking at the fragments alone.

The antecedent-task was handled more satisfactorily, as Table 4 shows. For this task, a naïve baseline ("always take previous utterance") preforms relatively well already; however, all classifiers were able to improve on this, with a slight advantage for the maxent model ($f(0.5) = 0.76$). As the entry for MaxEnt shows, adding to the baseline-features

Data Set	Cl.	Recall	Precision	$f(0.5)$	$f(1.0)$	$f(2.0)$
B; bl	m	0.00	0.00	0.00	0.00	0.00
B; bl+nrb+bft	m	36.39	31.16	0.31	0.33	0.35
B; all	m	**40.61**	**44.10**	**0.43**	**0.42**	**0.41**
UB; all	m	**22.13**	**65.06**	**0.47**	**0.33**	**0.25**
B; bl	t	31.77	21.20	0.22	0.24	0.28
B; bl+nrb+bpr+bds	t	**42.18**	**41.26**	**0.41**	**0.42**	**0.42**
B; all	t	44.54	32.74	0.34	0.37	0.41
UB; bl+nrb	t	**26.22**	**59.05**	**0.47**	**0.36**	**0.29**
B; bl	s	21.07	16.95	0.17	0.18	0.20
B; bl+nrb+bft+bds	s	**36.37**	**49.28**	**0.46**	**0.41**	**0.38**
B; all	s	36.67	43.31	0.42	0.40	0.38
UB; bl+nrb	s	**28.28**	**57.88**	**0.48**	**0.38**	**0.31**
B	s.s	32.57	42.96	0.40	0.36	0.34
B	b	55.62	19.75	0.23	0.29	0.41
UB	b	66.50	20.00	0.23	0.31	0.45

Table 3: Results for the fragment task. (Cl. = classifier used, where s = slipper, s.s = slipper + set-valued features, t = timbl, m = maxent, b = naive bayes; UB/B = (un)balanced training data.)

Data Set	Cl.	Recall	Precision	$f(0.5)$	$f(1.0)$	$f(2.0)$
dis=1	-	44.95	44.81	0.45	0.45	0.45
UB; bl	m	0	0	0.0	0.0	0.0
UB; bl+awh	m	43.21	52.90	0.50	0.47	0.45
UB; bl+awh+god	m	36.98	75.31	0.62	0.50	0.41
UB; bl+awh+god+lbe+lal+iqu+nra+buh	m	**64.26**	**80.39**	**0.76**	**0.71**	**0.67**
UB; all	m	58.16	73.57	0.69	0.64	0.60
UB; bl	s	0.00	0.00	0.00	0.00	0.00
UB; bl+aqm	s	36.65	78.44	0.63	0.49	0.41
UB; bl+aqm+rab+iqu+lal	s	**49.72**	**79.75**	**0.71**	**0.61**	**0.54**
UB; all	s	49.43	72.57	0.66	0.58	0.52
UB; bl	t	0	0	0.0	0.0	0.0
UB; bl+aqm	t	36.98	73.58	0.61	0.49	0.41
UB; bl+aqm+awh+rab+iqu	t	**46.41**	**77.65**	**0.68**	**0.58**	**0.50**
UB; all	t	60.57	58.74	0.59	0.60	0.60

Table 4: Results for the antecedent task.

Data Set	Cl.	Recall	Precision	$f(0.5)$	$f(1.0)$	$f(2.0)$
B; bl	m	0.00	0.00	0.00	0.00	0.00
B; bl+rap	m	5.83	40.91	0.18	0.10	0.07
B; bl+rap+god	m	7.95	55.83	0.25	0.14	0.10
B; bl+rap+god+nspk	m	11.70	49.15	0.30	0.19	0.14
B; bl+rap+god+nspk+alt+awh+nra+lal	m	**20.27**	**50.02**	**0.38**	**0.28**	**0.23**
B; all	m	23.29	43.79	0.36	0.30	0.25
UB; bl+rap+god+nspk+iqu+nra+bds+rab+awh	m	**13.01**	**54.87**	**0.33**	**0.21**	**0.15**
B; bl	s	0.00	0.00	0.00	0.00	0.00
B; bl+god	s	11.80	35.60	0.25	0.17	0.13
B; bl+god+bds	s	14.44	46.98	0.32	0.22	0.17
B; all	s	**17.78**	**41.96**	**0.32**	**0.24**	**0.20**
UB; bl+alt+bds+god+sspk+rap	s	**11.37**	**56.34**	**0.31**	**0.19**	**0.13**
B; bl	t	0.00	0.00	0.00	0.00	0.00
B; bl+god	t	**17.20**	**29.09**	**0.25**	**0.21**	**0.19**
B; all	t	17.87	19.97	0.19	0.19	0.18
UB; bl+god+iqu+rab	t	**14.24**	**41.63**	**0.29**	**0.21**	**0.16**
B; bl+rab+buh	s.s	**8.63**	**54.20**	**0.26**	**0.15**	**0.10**

Table 5: Results for the combined task.

information about whether α is a question or not already boost the performance considerably. An analysis of the predictions of this model then indeed shows that it already captures cases of question and answer pairs quite well. Adding the similarity feature `god` then gives the model information about semantic relatedness, which, as hypothesised, captures elaboration-type relations (as in (1-b) and (1-c) above). Structural information (`iqu`) further improves the model; however, the remaining features only seem to add interfering information, for performance using the full feature-set is worse.

If one of the problems of the fragment-task was that information about the context is required to distinguish fragments and backchannels, then the hope could be that in the combined-task the classifier would able to capture these cases. However, the performance of all classifiers on this task is not satisfactory, as Table 5 shows; in fact, it is even slightly worse than the performance on the fragment task alone. We speculate that instead of of cancelling out mistakes in the other part of the task, the two goals (let β be a fragment, and α a typical antecedent) interfere during optimisation of the rules.

To summarise, we have shown that the task of identifying the antecedent of a given fragment is learnable, using a feature-set that combines structural and lexical features; in particular, the inclusion of a measure of semantic relatedness, which was computed via queries to an internet search engine, proved helpful. The task of identifying (*resolution-via-identity*) fragments, however, is hindered by the high number of non-sentential utterances which can be confused with the kinds of fragments we are interested in. Here it could be helpful to have a method that identifies and filters out backchannels, presumably using a much more local mechanism (as for example proposed in (Traum, 1994)). Similarly, the performance on the combined task is low, also due to a high number of confusions of backchannels and fragments. We discuss an alternative set-up below.

5 Related Work

To our knowledge, the tasks presented here have so far not been studied with a machine learning approach. The closest to our problem is (Fernández et al., 2004b), which discusses *classifying* certain types of fragments, namely questions of the type "Who?", "When?", etc. (*sluices*). However, that paper does not address the task of *identifying* those in a corpus (which in any case should be easier than our fragment-task, since those fragments cannot be confused with backchannels).

Overlapping from another direction is the work presented in (Zechner and Lavie, 2001), where the task of aligning questions and answers is tackled. This subsumes the task of identifying question-antecedents for short-answers, but again is presumably somewhat simpler than our general task, because questions are easier to identify. The authors also evaluate the use of the alignment of questions and answers in a summarisation system, and report an increase in summary fluency, without a compromise in informativeness. This is something we hope to be able to show for our tasks as well.

There are also similarities, especially of the antecedent task, to the pronoun resolution task (see e.g. (Strube and Müller, 2003; Poesio et al., 2004)). Interestingly, our results for the antecedent task are close to those reported for that task. The problem of identifying the units in need of an antecedent, however, is harder for us, due to the problem of there being a large number of non-sentential utterances that cannot be linked to a single utterance as antecedent. In general, this seems to be the main difference between our task and the ones mentioned here, which concentrate on more easily identified markables (questions, sluices, and pronouns).

6 Conclusions and Further Work

We have presented a machine learning approach to the task of identifying fragments and their antecedents in multi-party dialogue. This represents a well-defined subtask of computing discourse structure, which to our knowledge has not been studied so far. We have shown that the task of identifying the antecedent of a given fragment is learnable, using features that provide information about the structure of the discourse between antecedent and fragment, and about semantic closeness.

The other tasks, identifying fragments and the combined tasks, however, did not perform as well, mainly because of a high rate of confusions between general non-sentential utterances and frag-

ments (in our sense). In future work, we will try a modified approach, where the detection of fragments is integrated with a classification of utterances as backchannels, fragments, or full sentences, and where the antecedent task only ranks pairs, leaving open the possibility of excluding a supposed fragment by using contextual information. Lastly, we are planning to integrate our classifier into a processing pipeline after the pronoun resolution step, to see whether this would improve both our performance and the quality of automatic meeting summarisations.[9]

References

Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

William Cohen and Yoram Singer. 1999. A simple, fast, and effective rule learner. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, Orlando, Florida, July. AAAI.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2003. TiMBL: Tilburg memory based learner, version 5.0, reference guide. ILC Technical Report 03-10, Induction of Linguistic Knowledge; Tilburg University. Available from `http://ilk.uvt.nl/downloads/pub/...papers/ilk0310.pdf`.

Raquel Fernández and Jonathan Ginzburg. 2002. Non-sentential utterances in dialogue: A corpus-based study. In Kristiina Jokinen and Susan McRoy, editors, *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 15–26, Philadelphia, USA, July. ACL Special Interest Group on Dialog.

Raquel Fernández, Jonathan Ginzburg, Howard Gregory, and Shalom Lappin. 2004a. Shards: Fragment resolution in dialogue. In H. Bunt and R. Muskens, editors, *Computing Meaning*, volume 3. Kluwer.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2004b. Classifying ellipsis in dialogue: A machine learning approach. In *Proceedings of COLING 2004*, Geneva, Switzerland, August.

John S. Garofolo, Christophe D. Laprun, Martial Michel, Vincent M. Stanford, and Elham Tabassi. 2004. The NITS meeting room pilot corpus. In *Proceedings of the International Language Resources Conference (LREC04)*, Lisbon, Portugal, May.

J.J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and devlopment. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pages 517–520, San Francisco, USA, March.

Ron Kohavi and George H. John. 1997. Wrappers for feature selection. *Artificial Intelligence Journal*, 97(1–2):273–324.

Christoph Müller and Michael Strube. 2001. MMAX: A Tool for the Annotation of Multi-modal Corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 45–50, Seattle, USA, August.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics*, pages 144–151, Barcelona, Spain, July.

David Schlangen and Alex Lascarides. 2002. Resolving fragments using discourse information. In Johan Bos, Mary Ellen Foster, and Colin Matheson, editors, *Proceedings of the 6th International Workshop on Formal Semantics and Pragmatics of Dialogue (EDILOG 2002)*, pages 161–168, Edinburgh, September.

David Schlangen and Alex Lascarides. 2003. The interpretation of non-sentential utterances in dialogue. In Alexander Rudnicky, editor, *Proceedings of the 4th SIGdial workshop on Discourse and Dialogue*, Sapporo, Japan, July.

David Schlangen. 2003. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Lingustics*, Sapporo, Japan.

D. Traum and P. Heeman. 1997. Utterance units in spoken dialogue. In E. Maier, M. Mast, and S. LuperFoy, editors, *Dialogue Processing in Spoken Language Systems*, Lecture Notes in Artificial Intelligence. Springer-Verlag.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, Computer Science, University of Rochester, Rochester, USA, December.

Klaus Zechner and Anton Lavie. 2001. Increasing the coherence of spoken dialogue summaries by cross-speaker information linking. In *Proceedings of the NAAACL Workshop on Automatic Summarisation*, Pittsburgh, USA, June.

[9] **Acknowledgements:** We would like to acknowledge helpful discussions with Jason Baldridge and Michael Strube during the early stages of the project, and helpful comments from the anonymous reviewers.