Classifying Ellipsis in Dialogue: A Machine Learning Approach

Raquel FERNÁNDEZ, Jonathan GINZBURG and **Shalom LAPPIN**
Department of Computer Science
King's College London
Strand, London WC2R 2LS, UK
{raquel,ginzburg,lappin}@dcs.kcl.ac.uk

Abstract

This paper presents a machine learning approach to bare sluice disambiguation in dialogue. We extract a set of heuristic principles from a corpus-based sample and formulate them as probabilistic Horn clauses. We then use the predicates of such clauses to create a set of domain independent features to annotate an input dataset, and run two different machine learning algorithms: SLIPPER, a rule-based learning algorithm, and TiMBL, a memory-based system. Both learners perform well, yielding similar success rates of approx 90%. The results show that the features in terms of which we formulate our heuristic principles have significant predictive power, and that rules that closely resemble our Horn clauses can be learnt automatically from these features.

1 Introduction

The phenomenon of *sluicing*—bare *wh*-phrases that exhibit a sentential meaning—constitutes an empirically important construction which has been understudied from both theoretical and computational perspectives. Most theoretical analyses (e.g. (Ross, 1969; Chung et al., 1995)), focus on embedded sluices considered out of any dialogue context. They rarely look at *direct* sluices—sluices used in queries to request further elucidation of quantified parameters (e.g. (1a)). With a few isolated exceptions, these analyses also ignore a class of uses we refer to (following (Ginzburg and Sag, 2001) (G&S)) as *reprise* sluices. These are used to request clarification of reference of a constituent in a partially understood utterance, as in (1b).

(1) a. Cassie: I know someone who's a good kisser.
Catherine: Who? [KP4, 512][1]

b. Sue: You were getting a real panic then.
Angela: When? [KB6, 1888]

Our corpus investigation shows that the combined set of direct and reprise sluices constitutes more than 75% of all sluices in the British National Corpus (BNC). In fact, they make up approx. 33% of all *wh*-queries in the BNC.

In previous work (Fernández et al., to appear), we implemented G&S's analysis of direct sluices as part of an interpretation module in a dialogue system. In this paper we apply machine learning techniques to extract rules for sluice classification in dialogue.

In Section 2 we present our corpus study of classifying sluices into dialogue types and discuss the methodology we used in this study. Section 3 analyses the distribution patterns we identify and considers possible explanations for these patterns. In Section 4 we identify a number of heuristic principles for classifying each sluice dialogue type and formulate these principles as probability weighted Horn clauses. In Section 5, we then use the predicates of these clauses as features to annotate our corpus samples of sluices, and run two machine learning algorithms on these data sets. The first machine learner used, SLIPPER, extracts optimised rules for identifying sluice dialogue types that closely resemble our Horn clause principles. The second, TiMBL, uses a memory-based machine learning procedure to classify a sluice by generalising over similar environments in which the sluice occurs in a training set. Both algorithms performed well, yielding similar success rates of approximately 90%. This suggests that the features in terms of which we formulated our heuristic principles for classifying sluices were well motivated, and both learning algorithms that we used are well suited to the task of dialogue act classification for fragments on the basis of these features. We finally present our conclusions and future work in Section 6.

2 Corpus Study

2.1 The Corpus

Our corpus-based investigation of bare sluices has been performed using the ∼10 million word

[1] This notation indicates the British National Corpus file (KP4) and the sluice sentence number (512).

dialogue transcripts of the BNC. The corpus of bare sluices has been constructed using SCoRE (Purver, 2001), a tool that allows one to search the BNC using regular expressions.

The dialogue transcripts of the BNC contain 5183 bare sluices (i.e. 5183 sentences consisting of just a *wh*-word). We distinguish between the following classes of bare sluices: *what*, *who*, *when*, *where*, *why*, *how* and *which*. Given that only 15 bare *which* were found, we have also considered sluices of the form *which N*. Including *which N*, the corpus contains a total of 5343 sluices, whose distribution is shown in Table 1.

The annotation was performed on two different samples of sluices extracted from the total found in the dialogue transcripts of the BNC. The samples were created by arbitrarily selecting 50 sluices of each class (15 in the case of *which*). The first sample included all instances of bare *how* and bare *which* found, making up a total of 365 sluices. The second sample contained 50 instances of the remaining classes, making up a total of 300 sluices.

what	*why*	*who*	*where*
3045	1125	491	350
when	*which N*	*how*	*which*
107	160	50	15
Total: 5343			

Table 1: Total of sluices in the BNC

2.2 The Annotation Procedure

To classify the sluices in the first sample of our sub-corpus we used the categories described below. The classification was done by 3 expert annotators (the authors) independently.

Direct The utterer of the sluice understands the antecedent of the sluice without difficulty. The sluice queries for additional information that was explicitly or implicitly quantified away in the previous utterance.

(2) Caroline: I'm leaving this school.
Lyne: When? [KP3, 538]

Reprise The utterer of the sluice cannot understand some aspect of the previous utterance which the previous (or possibly not directly previous) speaker assumed as presupposed (typically a contextual parameter, except for *why*, where the relevant "parameter" is something like speaker intention or speaker justification).

(3) Geoffrey: What a useless fairy he was.
Susan: Who? [KCT, 1753]

Clarification The sluice is used to ask for clarification about the previous utterance as a whole.

(4) June: Only wanted a couple weeks.
Ada: What? [KB1, 3312]

Unclear It is difficult to understand what content the sluice conveys, possibly because the input is too poor to make a decision as to its resolution, as in the following example:

(5) Unknown : <unclear> <pause>
Josephine: Why? [KCN, 5007]

After annotating the first sample, we decided to add a new category to the above set. The sluices in the second sample were classified according to a set of five categories, including the following:

Wh-anaphor The antecedent of the sluice is a *wh*-phrase.

(6) Larna: We're gonna find poison apple and I know where that one is.
Charlotte: Where? [KD1, 2371]

2.3 Reliability

To evaluate the reliability of the annotation, we use the *kappa coefficient* (K) (Carletta, 1996), which measures pairwise agreement between a set of coders making category judgements, correcting for expected chance agreement. [2]

The agreement on the coding of the first sample of sluices was moderate ($K = 52$).[3] There were important differences amongst sluice classes: The lowest agreement was on the annotation for *why* ($K = 29$), *what* ($K = 32$) and *how* ($K = 32$), which suggests that these categories are highly ambiguous. Examination of the coincidence matrices shows that the largest confusions were between `reprise` and `clarification` in the case of *what*, and between `direct` and `reprise` for *why* and *how*. On the other hand, the agreement on classifying *who* was substantially higher ($K = 71$), with some disagreements between `direct` and `reprise`.

Agreement on the annotation of the 2nd sample was considerably higher although still not entirely convincing ($K = 61$). Overall agreement was improved in all classes, except for

[2] $K = P(A) - P(E)/1 - P(E)$, where P(A) is the proportion of actual agreements and P(E) is the proportion of expected agreement by chance, which depends on the number of relative frequencies of the categories under test. The denominator is the total proportion less the proportion of chance expectation.

[3] All values are shown as percentages.

where and *who*. Agreement on *what* improved slightly ($K = 39$), and it was substantially higher on *why* ($K = 52$), *when* ($K = 62$) and *which N* ($K = 64$).

Discussion Although the three coders may be considered *experts*, their training and familiarity with the data were not equal. This resulted in systematic differences in their annotations. Two of the coders (coder 1 and coder 2) had worked more extensively with the BNC dialogue transcripts and, crucially, with the definition of the categories to be applied. Leaving coder 3 out of the coder pool increases agreement very significantly: $K = 70$ in the first sample, and $K = 71$ in the second one. The agreement reached by the *more expert* pair of coders was high and stable. It provides a solid foundation for the current classification. It also indicates that it is not difficult to increase annotation agreement by relatively light training of coders.

3 Results: Distribution Patterns

In this section we report the results obtained from the corpus study described in Section 2. The study shows that the distribution of readings is significantly different for each class of sluice. Subsection 3.2 outlines a possible explanation of such distribution.

3.1 Sluice/Interpretation Correlations

The distribution of interpretations for each class of sluice is shown in Table 2. The distributions are presented as percentages of pairwise agreement (i.e. agreement between pairs of coders), leaving aside the `unclear` cases. This allows us to see the proportion made up by each interpretation for each sluice class, together with any correlations between sluice and interpretation. Distributions are similar over both samples, suggesting that corpus size is large enough to permit the identification of repeatable patterns.

Table 2 reveals interesting correlations between sluice classes and preferred interpretations. The most common interpretation for *what* is `clarification`, making up 69% in the first sample and 66% in the second one. *Why* sluices have a tendency to be `direct` (57%, 83%). The sluices with the highest probability of being `reprise` are *who* (76%, 95%), *which* (96%), *which N* (88%, 80%) and *where* (75%, 69%). On the other hand, *when* (67%, 65%) and *how* (87%) have a clear preference for `direct` interpretations.

	1st Sample			**2nd Sample**			
	`Dir`	`Rep`	`Cla`	`Dir`	`Rep`	`Cla`	`Wh-a`
what	9	22	69	7	23	66	4
why	57	43	0	83	14	0	3
who	24	76	0	0	95	0	5
where	25	75	0	22	69	0	9
when	67	33	0	65	29	0	6
which N	12	88	0	20	80	0	0
which	4	96	0	–	–	–	–
how	87	8	5	–	–	–	–

Table 2: Distributions as pairwise agr percentages

3.2 Explaining the Frequency Hierarchy

In order to gain a complete perspective on sluice distribution in the BNC, it is appropriate to combine the (averaged) percentages in Table 2 with the absolute number of sluices contained in the BNC (see Table 1), as displayed in Table 3:

what$_{cla}$	2040	whichN$_{rep}$	135
why$_{dir}$	775	when$_{dir}$	90
what$_{rep}$	670	who$_{dir}$	70
who$_{rep}$	410	where$_{dir}$	70
why$_{rep}$	345	how$_{dir}$	45
where$_{rep}$	250	when$_{rep}$	35
what$_{dir}$	240	whichN$_{dir}$	24

Table 3: Sluice Class Frequency - Estim. Tokens

For instance, although more than 70% of *why* sluices are `direct`, the absolute number of *why* sluices that are `reprise` exceeds the total number of *when* sluices by almost 3 to 1. Explicating the distribution in Table 3 is important in order to be able to understand among other issues whether we would expect a similar distribution to occur in a Spanish or Mandarin dialogue corpus; similarly, whether one would expect this distribution to be replicated across different domains. Here we restrict ourselves to sketching an explanation of a couple of striking patterns exhibited in Table 3.

One such pattern is the low frequency of *when* sluices, particularly by comparison with what one might expect to be its close cousin—*where*; indeed the `direct/reprise` splits are almost mirror images for *when* v. *where*. Another very notable pattern, alluded to above, is the high frequency of *why* sluices.[4]

The *when* v. *where* contrast provides one argument against (7), which is probably the null

[4] As we pointed out above, sluices are a common means of asking *wh*–interrogatives; in the case of *why*–interrogatives, this is even stronger—close to 50% of all such interrogatives in the BNC are sluices.

hypothesis w/r to the distribution of reprise sluices:

(7) **Frequency of antecedent hypothesis**: The frequency of a class of reprise sluices is directly correlated with the frequency of the class of its possible antecedents.

Clearly locative expressions do not outnumber temporal ones and certainly not by the proportion the data in Table 3 would require to maintain (7).[5] (Purver, 2004) provides additional data related to this—clarification requests of all types in the BNC that pertain to nominal antecedents outnumber such CRs that relate to verbal antecedents by 40:1, which does not correlate with the relative frequency of nominal v. verbal antecedents (about 1.3:1).

A more refined hypothesis, which at present we can only state quite informally, is (8):

(8) **Ease of grounding of antecedent hypothesis**: The frequency of a class of reprise sluices is directly correlated with the ease with which the class of its possible antecedents can be grounded (in the sense of (Clark, 1996; Traum, 1994)).

This latter hypothesis offers a route towards explaining the *when* v. *where* contrast. There are two factors at least which make grounding a temporal parameter significantly easier on the whole than grounding a locative parameter. The first factor is that conversationalists typically share a temporal ontology based on a clock and/or calendar. Although well structured locative ontologies do exist (e.g. grid points in a map), they are far less likely to be common currency. The natural ordering of clock/calendar-based ontologies reflected in grammatical devices such as *sequence of tense* is a second factor that favours temporal parameters over locatives.

From this perspective, the high frequency of *why* reprises is not surprising. Such reprises query either the justification for an antecedent assertion or the goal of an antecedent query. Speakers usually do not specify these explicitly. In fact, what requires explanation is why such reprises do not occur even more frequently than they actually do. To account for this, one has to appeal to considerations of the *importance* of anchoring a contextual parameter.[6]

A detailed explication of the distribution shown in Table 3 requires a detailed model of dialogue interaction. We have limited ourselves to suggesting that the distribution can be explicated on the basis of some quite general principles that regulate grounding.

4 Heuristics for sluice disambiguation

In this section we informally describe a set of heuristics for assigning an interpretation to bare sluices. In subsection 4.2, we show how our heuristics can be formalised as probabilistic sluice typing constraints.

4.1 Description of the heuristics

To maximise accuracy we have restricted ourselves to cases of three-way agreement among the three coders when considering the distribution patterns from which we obtained our heuristics. Looking at these patters we have arrived at the following general principles for resolving bare sluice types.

What The most likely interpretation is `clarification`. This seems to be the case when the antecedent utterance is a fragment, or when there is no linguistic antecedent. `Reprise` interpretations also provide a significant proportion (about 23%). If there is a pronoun (matching the appropriate semantic constraints) in the antecedent utterance, then the preferred interpretation is `reprise`:

(9) Andy: I don't know how to do it.
Nick: What? Garlic bread? [KPR, 1763]

Why The interpretation of *why* sluices tends to be `direct`. However, if the antecedent is a non-declarative utterance, or a negative declarative, the sluice is likely to be a `reprise`.

(10) Vicki: Were you buying this erm newspaper last week by any chance?
Frederick: Why? [KC3, 3388]

Who Sluices of this form show a very strong preference for `reprise` interpretation. In the majority of cases, the antecedent is either a proper name (11), or a personal pronoun.

[5] A rough estimate concerning the BNC can be extracted by counting the words that occur more than 1000 times. Of these approx 35k tokens are locative in nature and could serve as antecedents of *where*; the corresponding number for temporal expressions and *when* yields approx 80k tokens. These numbers are derived from a frequency list (Kilgarriff, 1998) of the demographic portion of the BNC.

[6] Another factor is the existence of default strategies for resolving such parameters, e.g. assuming that the question asked transparently expresses the querier's primary goal.

(11) Patrick: [...] then I realised that it was Fennite
Katherine: Who? [KCV, 4694]

Which/Which N Both sorts of sluices exhibit a strong tendency to `reprise`. In the overwhelming majority of reprise cases for both *which* and *which N*, the antecedent is a definite description like 'the button' in (12).

(12) Arthur: You press the button.
June: Which one? [KSS, 144]

Where The most likely interpretation of *where* sluices is `reprise`. In about 70% of the reprise cases, the antecedent of the sluice is a deictic locative pronoun like 'there' or 'here'. `Direct` interpretations are preferred when the antecedent utterance is declarative with no overt spatial location expression.

(13) Pat: You may find something in there actually.
Carole: Where? [KBH, 1817]

When If the antecedent utterance is a declarative and there is no time-denoting expression other than tense, the sluice will be interpreted as `direct`, as in example (14). On the other hand, deictic temporal expressions like 'then' trigger `reprise` interpretations.

(14) Caroline: I'm leaving this school.
Lyne: When? [KP3, 538]

How This class of sluice exhibits a very strong tendency to `direct` (87%). It appears that most of the antecedent utterances contain an accomplishment verb.

(15) Anthony: I've lost the, the whole work itself
Arthur: How? [KP1, 631]

4.2 Probabilistic Constraints

The problem we are addressing is typing of bare sluice tokens in dialogue. This problem is analogous to part-of-speech tagging, or to dialogue act classification.

We formulate our typing constraints as Horn clauses to achieve the most general and declarative expression of these conditions. The antecedent of a constraint uses predicates corresponding to dialogue relations, syntactic properties, and lexical content. The predicate of the consequent represents a sluice typing tag, which corresponds to a maximal type in the HPSG grammar that we used in implementing our dialogue system. Note that these constraints *cannot* be formulated at the level of the lexical entries of the *wh*-words since these distributions are specific to sluicing and not to non-elliptical *wh*-interrogatives.[7] As a first example, consider the following rule:

```
 sluice(x), where(x),
       ant_utt(y,x),
   contains(y,'there')   →   reprise(x) [.78]
```

This rule states that if `x` is a sluice construction with lexical head `where`, and its antecedent utterance (identified with the latest move in the dialogue) contains the word 'there', then `x` is a `reprise` sluice. Note that, as in a probabilistic context-free grammar (Booth, 1969), the rule is assigned a conditional probability. In the example above, .78 is the probability that the context described in the antecedent of the clause produces the interpretation specified in the consequent.[8]

The following three rules are concerned with the disambiguation of *why* sluice readings. The structure of the rules is the same as before. In this case however, the disambiguation is based on syntactic and semantic properties of the antecedent utterance as a whole (like polarity or mood), instead of focusing on a particular lexical item contained in such utterance.

```
        sluice(x), why(x),
ant_utt(y,x), non_decl(y)  →  reprise(x)  [.93]
        sluice(x), why(x),
ant_utt(y,x), pos_decl(y)  →  direct(x)   [.95]
        sluice(x), why(x),
ant_utt(y,x), neg_decl(y)  →  reprise(x)  [.40]
```

5 Applying Machine Learning

To evaluate our heuristics, we applied machine learning techniques to our corpus data. Our aim was to evaluate the predictive power of the features observed and to test whether the intuitive constraints formulated in the form of Horn clause rules could be learnt automatically from these features.

5.1 SLIPPER

We use a rule-based learning algorithm called SLIPPER (for Simple Learner with Iterative Pruning to Produce Error Reduction). SLIPPER (Cohen and Singer, 1999) combines the

[7] Thus, whereas Table 2 shows that approx. 70% of *who*-sluices are `reprise`, this is clearly not the case for non-elliptical *who*–interrogatives. For instance, the KB7 block in the BNC has 33 non-elliptical *who*–interrogatives. Of these at most 3 serve as reprise utterances.

[8] These probabilities have been extracted manually from the three-way agreement data.

separate-and-conquer approach used by most rule learners with confidence-rated boosting to create a compact rule set.

The output of SLIPPER is a weighted rule set, in which each rule is associated with a confidence level. The rule builder is used to find a rule set that separates each class from the remaining classes using growing and pruning techniques. To classify an instance x, one computes the sum of the confidences that cover x: if the sum is greater than zero, the positive class is predicted. For each class, the only rule with a negative confidence rating is a single default rule, which predicts membership in the remaining classes.

We decided to use SLIPPER for two main reasons: (1) it generates transparent, relatively compact rule sets that can provide interesting insights into the data, and (2) its if-then rules closely resemble our Horn clause constraints.

5.2 Experimental Setup

To generate the input data we took all three-way agreement instances plus those instances where there is agreement between coder 1 and coder 2, leaving out cases classified as `unclear`. We reclassified 9 instances in the first sample as `wh-anaphor`, and also included these data.[9] The total data set includes 351 datapoints. These were annotated according to the set of features shown in Table 4.

`sluice`	type of sluice
`mood`	mood of the antecedent utterance
`polarity`	polarity of the antecedent utterance
`frag`	whether the antecedent utterance is a fragment
`quant`	presence of a quantified expression
`deictic`	presence of a deictic pronoun
`proper_n`	presence of a proper name
`pro`	presence of a pronoun
`def_desc`	presence of a definite description
`wh`	presence of a *wh* word
`overt`	presence of any other potential antecedent expression

Table 4: Features

We use a total of 11 features. All features are nominal. Except for the `sluice` feature that indicates the sluice type, they are all boolean, i.e. they can take as value either `yes` or `no`. The features `mood`, `polarity` and `frag` refer to syntactic and semantic properties of the antecedent utterance as a whole. The remaining features, on the other hand, focus on a particular lexical item or construction contained in such utterance. They will take `yes` as a value if this element or construction exists *and*, it matches the semantic restrictions imposed by the sluice type. The feature `wh` will take a `yes` value only if there is a *wh*-word that is identical to the sluice type. Unknown or irrelevant values are indicated by a question mark. This allows us to express, for instance, that the presence of a proper name is irrelevant to determine the interpretation of a *where* sluice, while it is crucial when the sluice type is *who*. The feature `overt` takes `no` as value when there is no overt antecedent expression. It takes `yes` when there is an antecedent expression not captured by any other feature, and it is considered irrelevant (question mark value) when there is an antecedent expression defined by another feature.

5.3 Accuracy Results

We performed a 10-fold cross-validation on the total data set, obtaining an average success rate of 90.32%. Using leave-one-out cross-validation we obtained an average success rate of 84.05%. For the holdout method, we held over 100 instances as a testing data, and used the reminder (251 datapoints) for training. This yielded a success rate of 90%. Recall, precision and f-measure values are reported in Table 5.

category	**recall**	**precision**	**f-measure**
`direct`	96.67	85.29	90.62
`reprise`	88.89	94.12	91.43
`clarification`	83.33	71.44	76.92
`wh_anaphor`	80.00	100	88.89

Table 5: SLIPPER - Results

Using the holdout procedure, SLIPPER generated a set of 23 rules: 4 for `direct`, 13 for `reprise`, 1 for `clarification` and 1 for `wh-anaphor`, plus 4 default rules, one for each class. All features are used except for `frag`, which indicates that this feature does not play a significant role in determining the correct reading. The following rules are part of the rule set generated by SLIPPER:

```
direct not reprise|clarification|wh_anaphor :-
                 overt=no, polarity=pos (+1.06296)

reprise not direct|clarification|wh_anaphor :-
                              deictic=yes (+3.31703)

reprise not direct|clarification|wh_anaphor :-
                mood=non_decl, sluice=why (+1.66429)
```

[9] We reclassified those instances that had motivated the introduction of the `wh-anaphor` category for the second sample. Given that there were no disagreements involving this category, such reclassification was straightforward.

5.4 Comparing SLIPPER and TiMBL

Although SLIPPER seems to be especially well suited for the task at hand, we decided to run a different learning algorithm on the same training and testing data sets and compare the results obtained. For this experiment we used TiMBL, a memory-based learning algorithm developed at Tilburg University (Daelemans et al., 2003). As with all memory-based machine learners, TiMBL stores representations of instances from the training set explicitly in memory. In the prediction phase, the similarity between a new test instance and all examples in memory is computed using some distance metric. The system will assign the most frequent category within the set of most similar examples (the k-nearest neighbours). As a distance metric we used information-gain feature weighting, which weights each feature according to the amount of information it contributes to the correct class label.

The results obtained are very similar to the previous ones. TiMBL yields a success rate of 89%. Recall, precision and f-measure values are shown in Table 6. As expected, the feature that received a lowest weighting was `frag`.

category	recall	precision	f-measure
`direct`	86.60	86.60	86.6
`reprise`	88.89	90.50	89.68
`clarification`	83.33	71.44	76.92
`wh_anaphor`	100	100	100

Table 6: TiMBL -Results

6 Conclusion and Further Work

In this paper we have presented a machine learning approach to bare sluice classification in dialogue using corpus-based empirical data. From these data, we have extracted a set of heuristic principles for sluice disambiguation and formulated such principles as probability weighted Horn clauses. We have then used the predicates of these clauses as features to annotate an input dataset, and ran two different machine learning algorithms: SLIPPER, a rule-based learning algorithm, and TiMBL, a memory-based learning system. SLIPPER has the advantage of generating transparent rules that closely resemble our Horn clause constraints. Both algorithms, however, perform well, yielding to similar success rates of approximately 90%. This shows that the features we used to formulate our heuristic principles were well motivated, except perhaps for the feature `frag`, which does not seem to have a significant predictive power. The two algorithms we used seem to be well suited to the task of sluice classification in dialogue on the basis of these features.

In the future we will attempt to construct an automatic procedure for annotating a dialogue corpus with the features presented here, to which both machine learning algorithms apply.

References

T. Booth. 1969. Probabilistic representation of formal languages. In *IEEE Conference Record of the 1969 Tenth Annual Symposium of Switching and Automata Theory.*

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 2(22):249–255.

S. Chung, W. Ladusaw, and J. McCloskey. 1995. Sluicing and logical form. *Natural Language Semantics*, 3:239–282.

H. H. Clark. 1996. *Using Language.* Cambridge University Press, Cambridge.

W. Cohen and Y. Singer. 1999. A simple, fast, and effective rule learner. In *Proc. of the 16th National Conference on AI.*

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. 2003. TiMBL: Tilburg Memory Based Learner, Reference Guide. Technical Report ILK-0310, U. of Tilburg.

R. Fernández, J. Ginzburg, H. Gregory, and S. Lappin. (to appear). SHARDS: Fragment resolution in dialogue. In H. Bunt and R. Muskens, editors, *Computing Meaning*, volume 3. Kluwer.

J. Ginzburg and I. Sag. 2001. *Interrogative Investigations.* CSLI Publications, Stanford, California.

A. Kilgarriff. 1998. BNC Database and Word Frequency Lists. www.itri.bton.ac.uk/ ~Adam.Kilgarriff/ bnc-readme.html.

M. Purver. 2001. SCoRE: A tool for searching the BNC. Technical Report TR-01-07, Dept. of Computer Science, King's College London.

M. Purver. 2004. *The Theory and Use of Clarification in Dialogue.* Ph.D. thesis, King's College, London, forthcoming.

J. Ross. 1969. Guess who. In *Proc. of the 5th annual Meeting of the Chicago Linguistics Society*, pages 252–286, Chicago. CLS.

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