Proceedings of the Third SIGdial Workshop on Discourse and Dialogue,
Philadelphia, July 2002, pp. 174-183. Association for Computational Linguistics.

Processing Unknown Words in a Dialogue System

Matthew Purver
Department of Computer Science
King's College London
Strand, London WC2R 2LS, UK
matthew.purver@kcl.ac.uk

Abstract

This paper describes a method of processing unknown words in a HPSG-based dialogue system, with acquisition of lexical semantics via clarification questions answered by the user. Use of a highly contextualized semantic representation, together with an utterance-anaphoric view of clarification, allows the clarificational dialogue to be integrated within the grammar and governed by standard rules of conversation.

1 Introduction

Most natural language processing applications have to deal with unknown words – as the lexicon of any natural language is infinite and ever-expanding, their appearance is practically guaranteed. Systems which depend on detailed lexical syntactic and semantic information must somehow acquire this information for unknown words.

Techniques for automatic acquisition of words depend either on world/domain knowledge (as in script-based approaches), or plentiful contextual information (as in corpus-based approaches). In an open- or wide-domain dialogue system, neither of these are available. This paper describes a solution to this problem based on interaction with the user via clarification questions. This approach is entirely integrated within a HPSG grammar and an information state-based dialogue move engine, and has been implemented in Prolog within a text-based research dialogue system.

In the next two sections I give some background: firstly on approaches to the processing of unknown words, and secondly on a theory of clarification questions. Section 4 then gives an overview of the proposed method. This is then described in detail in sections 5 (initial processing), 6 (clarification) and 7 (integration). Some tentative indications as to coverage are given in section 8, further proposed work is briefly introduced in section 9, and conclusions are drawn in section 10.

2 Background – Unknown Words

Approaches to unknown word processing can be broadly divided into two categories: those which depend on user intervention and those which do not. The latter have received more attention in recent years, as automatic operation is desirable in large-scale applications such as text summarization or classification, although the former can be acceptable in applications where user interaction is the norm, such as translation and dialogue systems.

Automatic systems can again be divided into two classes: those which use predefined world/domain knowledge to infer properties of the unknown word, and those which are experience-based, using corpus data or context. Knowledge-based systems (Granger, 1977; Russell, 1993) use scripted information about a known situation which allow semantic information about a word to be inferred. In this way detailed information can be gained about words which play major roles in sentences, although very little progress can be made with modifiers e.g. adjectives. More importantly, the domain is limited, and the approach cannot be applied to a system intended for open- or wide-domain use.

Experience-based approaches (Hastings, 1994; Pedersen, 1995; Thompson, 1998; Barg and Walther, 1998) avoid this requirement for domain knowledge and can acquire detailed syntactic information, but the semantic information gained tends to be limited to argument selection and broad category classification unless large amounts of data are available.

In a dialogue system, an unknown word must be processed on its first appearance – the only contextual information available is that given by the surrounding sentence. This may be sufficient to allow syntactic information to be inferred (at least sufficient to parse the sentence), but some user interaction will be required in order to acquire semantic information.

Previous approaches to user-guided acquisition often involve asking the user to select from a list of possible usages (Carter, 1992). Within a natural dialogue framework, however, clarificational dialogue must be used. This concept is not new: (Zernik, 1987) uses questions about the meaning of phrases in a system simulating a second language learner, and (Knight, 1996) proposes the use of questions about word meaning in a translation system. However, these systems understandably treat the clarification exchange as self-contained and governed by its own rules; but within a dialogue system it cannot necessarily be distinguished from the wider dialogue. This paper therefore outlines a method of integrating a clarificational approach within a grammar, making it a seamless part of the dialogue engine.

3 Background – Clarification Requests

The theory developed by (Ginzburg and Cooper, 2001; Ginzburg and Cooper, forthcoming) (G&C) provides an analysis of clarification questions together with a method of resolution of associated elliptical forms. This analysis is couched within a HPSG grammar and a Question-Under-Discussion (QUD) approach to dialogue context.

Clarification questions are regarded as being *utterance-anaphoric*, in that they refer to (a constituent of) a previous utterance (the utterance being clarified). This allows for, amongst other possibilities, an analysis of clarification questions in which their semantic content is a question concerning the content of such a constituent. In other words, the content of B's query *"Bo?"* in example (1) could be paraphrased as the question in example (2) (also shown as a typed feature structure in AVM [1]):

(1)
A: Did Bo leave?
B: **BO?**
A: Bo Smith.
B: Yes, half an hour ago.

(2) *"What is the intended content of your utterance "Bo"?"*

$$[1]\quad \begin{bmatrix} \textit{question} & \\ \text{PARAMS} & \left\{\boxed{2}\right\} \\ \text{PROP} \mid \text{SOA} & \begin{bmatrix} \textit{spkr-meaning-rel} & \\ \text{SPKR} & \boxed{a} \\ \text{SIGN} & \begin{bmatrix} \textit{sign} & \\ \text{PHON} & \langle \text{bo} \rangle \\ \text{CONTENT} & \boxed{1}\begin{bmatrix} \textit{param} \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \boxed{2} \end{bmatrix} \end{bmatrix}$$

To allow access to constituents (in order to build these questions), their version of HPSG assumes a feature CONSTITS whose value is a set of all constituents of a given sign – this addition is assumed here.

G&C also posit a modified semantic representation within HPSG which allows an analysis of how clarification questions arise. Standard versions of HPSG directly encode idealized content (that which an idealized agent would associate with a sign) within the value for the CONTENT feature. Instead, they propose a representation which allows contextual dependence: contextually dependent parameters are abstracted to a set which is the value of a new C-PARAMS feature – see AVM [2].[1] This allows the sign to

[1] The C-PARAMS feature also includes contextual information such as the identities of speaker and addressee,

be viewed as a λ-abstract, or a *meaning* in the Montogovian sense.

[2]
$$\begin{bmatrix} \text{C-PARAMS} & \left\{\begin{bmatrix}\text{INDEX} & \boxed{1}\\ \text{RESTR} & \left\{named(\boxed{1},Bo)\right\}\end{bmatrix}\right\} \\ \text{CONTENT} & \begin{bmatrix} \textit{ask-rel} & \\ \text{ASKER} & \boxed{a} \\ \text{ASKED} & \boxed{b} \\ \text{MSG-ARG} & \begin{bmatrix} \textit{question} & \\ \text{PARAMS} & \{\} \\ \text{PROP} \mid \text{SOA} & \begin{bmatrix} \textit{leave-rel} & \\ \text{AGENT} & \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

$$\lambda\{X\}[named(Bo,X)]ask(A,B,\langle\langle leave(X)\rangle\rangle)$$

The grounding process for an addressee can then be thought of as a process of establishing the referents of these parameters; any failure so to do will result in the formation of a clarification question with the purpose of querying the relevant parameter.

G&C also assume a dialogue information state which includes memory for *utterances* rather than solely their content (e.g. *dialogue moves*) – thus providing referents for an utterance-anaphoric approach. This is also assumed here.

4 Overview

The intention of this paper is to outline an approach to unknown word processing that can be used in an open-domain dialogue system, and which mirrors human-human dialogues such as example (3)[2] to produce natural dialogues such as the imaginary example (4). The approach taken is thus necessarily one involving clarification dialogue. As such it has similarities to the approach of (Knight, 1996). The approach presented here, however, is based upon a full grammatical analysis of clarification questions and their answers, and as such can be integrated within the dialogue system grammar.

(3)[3]
Ruth: Wouldn't argue with that one **iota**.
Paul: **What does iota mean?**
Ruth: **Scrap.**
Paul: One what?
Ruth: Scrap.

(4)
Usr: I would like to travel by **pullman**.
Sys: (What do you mean by) **Pullman?**
Usr: (I mean) **First class train**.
Sys: OK. I'm afraid first class is fully booked.

The approach taken follows the algorithm shown in listing 1. An utterance containing a unknown word is parsed using an existing grammar. During the grounding process, interpretation of semantic parameters corresponding to any unknown words will fail. This causes a clarification question to be produced. The answer to this question is then used to resolve the semantics of the unknown word. After this resolution, the word can be added to the system lexicon.

```
1. For new utterance U:
   parse to give sign S, add to PENDING stack.
2. Attempt to ground S.
3. If successful, remove S from PENDING,
   add S to LATEST-MOVE and skip to end.
4. Otherwise form clarification question Q
   about ungrounded constituent C of S.
5. Ask Q.
6. Process answer A as usual
   (i.e. starting from step 1).
7. If A answers Q:
   (a) Add new lexical entry for C.
   (b) Return to step 2.
8. Otherwise return to step 4.
```

Listing 1: Algorithm for Lexical Acquisition

Before this process can be accomplished as described, several steps are required. Firstly, the grammar used must be capable of parsing (giving a suitable syntactic and semantic analysis) sentences with unknown words. This must involve a treatment of such words that allows their meaning to be abstracted in a similar way to the

and time of utterance – for simplification, these are not shown here but should be assumed.

[2] Examples given here are taken from the British National Corpus (BNC) (Burnard, 2000) and were found using SCoRE (Purver, 2001).

[3] BNC file KD0, sentences 944–948

treatment of names described by G&C, and this is described in section 5.

Secondly, the clarification question must be formed, raised and answered. Again, this requires some modifications to the theory of G&C, and requires a method of answer resolution. This is described in section 6.

Lastly, we need some method of using the answering of the clarification question to resolve the unknown word semantics and form a new lexical entry. This is described in section 7.

5 Initial Processing

5.1 Syntax

Parsing can be achieved by allowing words not in the lexicon to be represented as a disjunction of generic entries for all open-class syntactic categories, following (Erbach, 1990). The entries are generic in that only syntactic selectional restrictions and semantic predicate-argument structure are included, while the lexical semantics are left underspecified (see below).

Examination of the final state of the parser will determine the correct syntactic category chosen. At this stage, ambiguity is possible (in the sentence *"I saw her X"*, *X* could be a noun or a verb). This ambiguity could be reduced by part-of-speech (PoS) tagging based on orthographic form prior to parsing,[4] and will be reduced further during clarification (see section 6.3.2).

5.2 Semantics

The semantic analysis must give us the capability to ground words in context (and thus fail to do so in the case of unknown words, or indeed words that are in the lexicon but whose relevance or reference cannot be understood). The analysis of G&C gives us a starting point for this, but (as they point out) the contextualized representation must be extended to allow relation names to be added to the C-PARAMS set.

5.2.1 Nouns

This modification is relatively simple for nouns: the *parameter* which makes up the semantic content of the noun is abstracted directly into the C-PARAMS set in the same way as is described for proper names by G&C. An example is shown in AVM [3] below. In order to remain independent of any particular theory of lexical semantics, a simple LEXSEM attribute is shown here which is taken to be a semantic relation name.[5]

For unknown words, the content must be underspecified: we stipulate only that its RESTR set must have at least one member specifying a lexical semantic relation. This relation is left uninstantiated (i.e. the LEXSEM feature value is a place-holder variable), as shown in AVM [4].

$$[3]\quad \begin{bmatrix} \text{PHON} & \langle \text{train} \rangle \\ \text{CONTENT} & \boxed{1} \\ \text{C-PARAMS} & \left\{ \boxed{1} \begin{bmatrix} \textit{parameter} & \\ \text{INDEX} & \boxed{i} \\ \text{RESTR} & \left\{ \begin{bmatrix} \textit{noun-rel} & \\ \text{LEXSEM} & \text{train} \\ \text{INSTANCE} & \boxed{i} \end{bmatrix} \right\} \end{bmatrix} \right\} \end{bmatrix}$$

$$[4]\quad \begin{bmatrix} \text{PHON} & \langle \text{pullman} \rangle \\ \text{CONTENT} & \boxed{1} \\ \text{C-PARAMS} & \left\{ \boxed{1} \begin{bmatrix} \textit{parameter} & \\ \text{INDEX} & \boxed{i} \\ \text{RESTR} & \left\{ \ldots \begin{bmatrix} \textit{noun-rel} & \\ \text{LEXSEM} & \text{X} \\ \text{INSTANCE} & \boxed{i} \end{bmatrix} \ldots \right\} \end{bmatrix} \right\} \end{bmatrix}$$

5.2.2 Verbs

For verbs, a further modification is required. In the standard HPSG theory, the content of a headed construction is inherited from the head daughter – in the case of sentences this head daughter is the verb, and thus the content of the verb includes the predicate-argument structure of the sentence. However, this information is not required in the C-PARAMS set.

In consequence, an event-based representation is suggested (see e.g. (Copestake et al., 1999)),

[4] PoS-tagging could also increase parse efficiency by reducing the number of alternative lexical entries considered. It is possible to incorporate PoS-tagging within a stochastic HPSG grammar by associating probabilities with features that connect PHON and HEAD attributes. A version of this approach is currently being implemented using probabilities from MXPOST (Ratnaparkhi, 1996), but its trainability has yet to be assessed.

[5] A treatment of adjectives will also follow these lines, although this has not yet been implemented.

wherein a parameter describing the event contains that part of the semantic content concerned with the nature of the verb relation (but not the predicate-argument structure). This allows this part of the content to be abstracted, leaving predicate-argument role information as part of CONTENT while allowing all lexical semantic information to be present in C-PARAMS:

[5]
$$\begin{bmatrix} \text{CONTENT} & \begin{bmatrix} \textit{monadic-verb-rel} & \\ \text{AGENT} & \boxed{u} \\ \text{EVENT} & \boxed{e} \\ \text{EV-DESC} & \boxed{1} \end{bmatrix} \\ \text{C-PARAMS} & \left\{ \boxed{1} \begin{bmatrix} \textit{parameter} & \\ \text{INDEX} & \boxed{e} \\ \text{RESTR} & \left\{ \begin{bmatrix} \textit{event-rel} & \\ \text{LEXSEM} & \text{snore} \\ \text{INSTANCE} & \boxed{e} \end{bmatrix} \right\} \end{bmatrix} \right\} \end{bmatrix}$$

This event-based representation is not the only possible solution – in particular, the relational semantic parameter need not be considered to be describing an *event*: the crucial factor as far as this discussion is concerned is only that it is contained in an attribute within CONTENT, allowing it to be identified with a member of C-PARAMS.[6]

This separation of lexical semantics from argument structure allows the former to be underspecified (and to be the subject of clarification) while the latter is determined (at least to some extent) by parsing.[7] This avoids much of the ambiguity of semantic structure determination described by (Knight, 1996).

6 Clarification

6.1 Grounding

The representation built therefore takes a form as shown in AVM [6] for the sentence *"I would like to travel by pullman"*:

[6]
$$\begin{bmatrix} \text{PHON} & \langle \text{i, want, to, travel, by, pullman} \rangle \\ \text{CONTENT} & \begin{bmatrix} \text{AGENT} & \boxed{u} \\ \text{METHOD} & \boxed{i} \\ \text{EVENT} & \boxed{e} \\ \text{EV-DESC} & \boxed{1} \end{bmatrix} \\ \text{C-PARAMS} & \left\{ \begin{array}{l} \boxed{1} \begin{bmatrix} \text{INDEX} & \boxed{e} \\ \text{RESTR} & \left\{ \begin{bmatrix} \text{LEXSEM} & \text{travel} \\ \text{INSTANCE} & \boxed{e} \end{bmatrix} \right\} \end{bmatrix}, \\ \begin{bmatrix} \text{INDEX} & \boxed{u} \\ \text{RESTR} & \left\{ \begin{bmatrix} \text{NAME} & \text{user} \\ \text{INSTANCE} & \boxed{u} \end{bmatrix} \right\} \end{bmatrix}, \\ \boxed{2} \begin{bmatrix} \text{INDEX} & \boxed{i} \\ \text{RESTR} & \left\{ \ldots \begin{bmatrix} \text{LEXSEM} & \text{X} \\ \text{INSTANCE} & \boxed{i} \end{bmatrix} \ldots \right\} \end{bmatrix} \end{array} \right\} \\ \text{CONSTITS} & \left\{ \ldots \begin{bmatrix} \text{PHON} & \langle \text{pullman} \rangle \\ \text{CONTENT} & \boxed{2} \\ \text{C-PARAMS} & \left\{ \boxed{2} \right\} \end{bmatrix} \ldots \right\} \end{bmatrix}$$

The grounding process now consists of anchoring the members of the C-PARAMS set in context. For proper names & pronouns such as *I*, this will involve finding a referent for which the specified restriction (in this case `user`) is satisfied, or possibly accommodating such a referent into the context. For lexical relations such as `travel` & `pullman`, it will involve finding a conceptual entry corresponding to the relation specified. With unknown words, where the relation semantics is underspecified, this process must fail.[8]

6.2 Questioning

Formation of a grammatically analyzable clarification question can now proceed almost as proposed by G&C, although some modifications are necessary to ensure that the question is correct for verbs – again, the desired question is not a question about the CONTENT of the verb (which is, by extension, the content of the whole sentence), but merely that part of the content available in C-PARAMS. For nouns, names etc. this modification makes no difference as the entire content is contained within C-PARAMS.

G&C's *parameter identification* context coercion operation, which allows the unknown constituent of the previous utterance to be made

[6] A treatment of adverbs will also follow these lines, although this has not yet been implemented.

[7] For example, a word determined by the parser to be a subject raising verb will have a generic "subject raising" argument structure – only the lexical semantics are underspecified.

[8] Given a suitable grounding process, failure could also be ensured when in-vocabulary words are used with a sense that cannot be understood in context – see section 9.

contextually available, now needs to involve question formation by abstraction not of the value of CONTENT but the member(s)[9] of C-PARAMS, as shown here:

$$\begin{bmatrix} \textit{root-cl} & \\ \text{C-PARAMS} & \left\{\ldots\boxed{1}\ldots\right\} \\ \text{CONSTITS} & \left\{\ldots\boxed{2}\begin{bmatrix}\text{C-PARAMS} & \boxed{4}\left\{\boxed{1}\right\}\end{bmatrix}\ldots\right\} \end{bmatrix}$$

$$\Rightarrow$$

$$\begin{bmatrix} \text{CONTENT} \mid \text{MSG-ARG} & \begin{bmatrix}\textit{question} & \\ \text{PROP} & \boxed{3}\end{bmatrix} \\ \text{SAL-UTT} & \boxed{2} \\ \text{MAX-QUD} & \begin{bmatrix} \textit{question} & \\ \text{PARAMS} & \boxed{4} \\ \text{PROP} & \boxed{3}\begin{bmatrix}\text{SOA} & \begin{bmatrix}\textit{spkr-mean-rel} & \\ \text{SPKR} & \boxed{u} \\ \text{SIGN} & \boxed{2} \\ \text{PARAMS} & \boxed{4}\end{bmatrix}\end{bmatrix} \end{bmatrix} \end{bmatrix}$$

This makes the unknown constituent contextually salient (as SAL-UTT), and makes the maximal question-under-discussion (MAX-QUD) the desired clarification question *"What is the intended meaning of your utterance X?"*[10]

Removing G&C's syntactic restriction (to NPs) on their *utterance-anaphoric-phrase* type then allows any unknown constituent to be clarified. The analysis that would be obtained for an elliptical clarification question *"Pullman?"* is shown here, although non-elliptical versions such as *"What do you mean by 'pullman'?"* are also available (and their use might be preferred due to the reduced ambiguity[11]).

$$\text{[7]}\quad \begin{bmatrix} \text{PHON} & \boxed{1}\left\langle \text{pullman} \right\rangle \\ \text{CONTENT} \mid \text{MSG-ARG} & \boxed{5} \\ \text{SAL-UTT} & \boxed{2}\begin{bmatrix}\text{PHON} & \boxed{1} \\ \text{C-PARAMS} & \left\{\begin{bmatrix}\text{INDEX} & \boxed{i} \\ \text{RESTR} & \left\{\ldots\begin{bmatrix}\text{LEXSEM} & \text{X} \\ \text{INSTANCE} & \boxed{i}\end{bmatrix}\ldots\right\}\end{bmatrix}\right\}\end{bmatrix} \\ \text{MAX-QUD} & \boxed{5}\begin{bmatrix}\textit{question} & \\ \text{PARAMS} & \boxed{4} \\ \text{PROP} \mid \text{SOA} & \begin{bmatrix}\textit{spkr-mean-rel} & \\ \text{SPKR} & \boxed{u} \\ \text{SIGN} & \boxed{2}\begin{bmatrix}\text{C-PARAMS} & \boxed{4}\end{bmatrix} \\ \text{PARAMS} & \boxed{4}\end{bmatrix}\end{bmatrix} \end{bmatrix}$$

The context coercion operation and question formation can be expressed as an information state update rule, shown in listing 2 in the TrindiKit notation (Traum et al., 1999):

```
rule( clarifyUsrUnknownConstit,
  [ fst#rec( shared^lu^utt, Utt ),
    sem :: unknown_constit( Utt, Constit )
  ],
  [ sem :: clarification_qustn( Constit, Q ),
    push#rec( shared^qud, Q ),
    push#rec( private^agenda, findout( Q ) )
  ] ).
```

Listing 2: Coercion Operation Update Rule

6.3 Answering

Once this clarification question has been asked, any forthcoming suitable answer must be processed. If such an answer takes a full non-elliptical form such as *"By my utterance 'pullman' I meant 'first class train"'*, the treatment within the grammar of *"mean"* as taking an utterance as argument allows it to be parsed and interpreted as an answer directly.[12]

Such answers do occur – see example (5). However, a much more common scenario is one in which the answer is elliptical – as in *"(Oh,*

[9] For the unknown words under consideration here, C-PARAMS will always be a singleton set whose single member corresponds to the lexical semantics of the word. If this analysis is to be extended to e.g. phrases, a multiple query could be formed by abstracting all the parameters in C-PARAMS, but the question of whether the predicate-argument structure component of CONTENT can also be queried may have to be addressed.

[10] It should be noted that this is not the same as *"What is a X?"* This avoids a problem which could arise when querying words that can have different meanings in different contexts. When clarifying the use *in context* of the word *cat*, we are not expecting an exhaustive answer giving all the possible meanings of *cat* (*feline/person-/boat/whip/vehicle/etc.*), but the meaning intended by the speaker in the particular utterance being clarified.

[11] This point was stressed by a SIGdial reviewer, who pointed out that elliptical questions can be misinterpreted as yes-no "check" questions. Indeed, the BNC shows several examples of such misinterpretation.

[12] At least two lexical entries for *mean* are required, and both require arguments to be inferred anaphorically from context: *"I mean 'first class train"'* requires the assumption of an utterance argument (*"**By the utterance we are discussing** I mean ..."*), while *"'Pullman' means 'first class train"'* requires the assumption of a speaker argument (*" ... **when used by me/anyone**".*)

you know,) first class train" – see examples (3) and (6).

In this case some method of ellipsis resolution is required to assign the full semantic content to the answer. Here a QUD-based method similar to that used by the SHARDS system (Gregory, 2001) can be used.[13] However, slight modifications are necessary both to handle verbs correctly and to adapt to the C-PARAMS approach.

(5)[14]
DW: many of which are just the sort of things we would expect to be in the **prebiotic** soup.
TB: **What do you mean by prebiotic?**
DW: **Prebiotic means** erm a system whereby biological processes have not actually started

(6)[15]
Catriona: **What does squire mean?** <pause>
Father: **Esquire.**
Catriona: Oh.

6.3.1 Constituent Answers

The usual definition of the *declarative-fragment-clause* type unifies only the INDEX values of the elliptical answer and the salient utterance in the question. In this case we desire the whole parameter to be unified, as the RESTR feature contains the semantic relation information.

In addition, the treatment is extended to verbs by removing any syntactic constraint on the head daughter, and by using the C-PARAMS value instead of CONTENT, as already explained above.[16]

The resulting modified definition is shown (simplified here) in AVM [8]:

$$[8]\quad \begin{bmatrix} \textit{decl-frag-cl} & \\ \text{CONTENT} & \boxed{3} \\ \text{HEAD-DTR} & \begin{bmatrix} \text{CATEGORY} & \boxed{1} \\ \text{C-PARAMS} & \boxed{2} \end{bmatrix} \\ \text{SAL-UTT} & \begin{bmatrix} \text{CATEGORY} & \boxed{1} \\ \text{C-PARAMS} & \boxed{2} \end{bmatrix} \\ \text{MAX-QUD} & \begin{bmatrix} \text{PARAMS} & \boxed{2} \\ \text{PROP} & \boxed{3} \end{bmatrix} \end{bmatrix}$$

6.3.2 Resolution

Using this phrase type, an elliptical fragment answer *"first class train"* will be resolved as shown in AVM [9] below.

$$[9]\quad \begin{bmatrix} \text{CONTENT} \mid \text{MSG-ARG}\ \boxed{3} & \begin{bmatrix} \text{SOA} & \begin{bmatrix} \textit{spkr-mean-rel} & \\ \text{SPKR} & \boxed{u} \\ \text{SIGN} & \boxed{4}\begin{bmatrix} \text{C-PARAMS} & \boxed{2} \end{bmatrix} \\ \text{PARAMS} & \boxed{1} \end{bmatrix} \end{bmatrix} \\ \text{HEAD-DTR} & \begin{bmatrix} \text{PHON} & \left\langle \text{first, class, train} \right\rangle \\ \text{CATEGORY} & \boxed{1} \\ \text{C-PARAMS}\boxed{2} & \left\{ \begin{bmatrix} \text{INDEX} & \boxed{j} \\ \text{RESTR} & \left\{ \begin{bmatrix} \text{LEXSEM} & \text{train} \\ \text{INSTANCE} & \boxed{j} \end{bmatrix}, \begin{bmatrix} \text{LEXSEM} & \text{first\_cl} \\ \text{INSTANCE} & \boxed{j} \end{bmatrix} \right\} \end{bmatrix} \right\} \end{bmatrix} \\ \text{SAL-UTT}\ \boxed{4} & \begin{bmatrix} \text{PHON} & \left\langle \text{pullman} \right\rangle \\ \text{CATEGORY} & \boxed{1} \\ \text{C-PARAMS}\boxed{2} & \left\{ \begin{bmatrix} \text{INDEX} & \boxed{i} \\ \text{RESTR} & \left\{ \ldots \begin{bmatrix} \text{LEXSEM} & \text{X} \\ \text{INSTANCE} & \boxed{i} \end{bmatrix} \ldots \right\} \end{bmatrix} \right\} \end{bmatrix} \\ \text{MAX-QUD} & \begin{bmatrix} \textit{question} & \\ \text{PARAMS} & \boxed{2} \\ \text{PROP} & \boxed{3} \end{bmatrix} \end{bmatrix}$$

The unification of C-PARAMS sets between HEAD-DTR (the answer) and SAL-UTT (the constituent being clarified) causes the lexical semantics of the unknown constituent to become instantiated as desired.

In addition, as the answer resolution enforces syntactic parallelism, any ambiguity of syntactic category will become resolved at this point.[17]

[13] In SHARDS, bare noun phrase fragments can be resolved as answers by referring to the Maximal Question Under Discussion (MAX-QUD – see e.g. (Ginzburg et al., 2001a)): given a question such as *"Who does John like?"*, the bare answer *"Mary"* can be resolved as meaning *"John likes Mary"*.

[14] BNC file KRH, sentences 2897–2899

[15] BNC file KP6, sentences 2006–2008

[16] This modification also allows resolution of bare verb fragment answers in general, providing an analysis for utterances such as B's in:

(7)
A: What are you doing tomorrow?
B: Going swimming.
(paraphrase: I am going swimming tomorrow)

[17] The requirement of strict syntactic parallelism may be too strong – corpus investigation of answers to clarification questions may shed some light on this.

7 Integration

By this stage the unknown word has been processed and an (underspecified) lexical entry determined during parsing. A question about its semantics has been formed, asked and answered. Until this point, all work has been performed by the grammar, a coercion operation to form a clarification question, and QUD-based ellipsis resolution.

All that is now needed is a dialogue rule that adds a new entry to the lexicon, corresponding to the newly instantiated sign, as shown in listing 3:

```
rule( integrateUsrAnswer,
    [ in#rec( shared^lu^moves, assert( P ) ),
      fst#rec( shared^qud, Q ),
      sem :: relevant_answer( Q, P ),
    ],
    [ pop#rec( shared^qud ),
      add#rec( shared^com, resolves( P, Q ) ),
      sem :: sal_utt( P, S ),
      lexicon :: add_entry( S )
    ] ).
```

Listing 3: Lexical Entry Addition Update Rule

If a more complex lexical semantic representation is being used, along the lines of the multi-dimensional approach of (Pustejovsky, 1998), it is likely that only certain dimensions have been instantiated by this unification, and the lexical entry is still underspecified in others. In this case, the word will now function as a known word in sentence contexts that pick out the newly known dimensions, and as an unknown word in contexts which select for those dimensions that remain underspecified.

8 Coverage

The approach outlined relies on direct answers to clarification questions being provided, both in order to recognise them as answers, and to allow the new information to be incorporated into the underspecified sign. A corpus investigation using the BNC to determine the proportion of direct answers given in human-human dialogue (and therefore to give some idea of likely coverage) has been attempted. Unfortunately the number of examples of clarification questions concerning unknown words in this corpus is small, so the results here must be regarded as preliminary at best.

Nevertheless they are encouraging: of answers to *"What does X mean?"* questions, in cases where X is an noun, 90% of answers were direct; for cases where X is an verb, 70% were direct.[18]

Interestingly all indirect examples occurred in "classroom test" situations, where the word being clarified was not unknown to the questioner – these situations tended to elicit examples of usage rather than direct answers.

Other corpora will be examined to find more examples in order to allow some confidence in these results – see below.

9 Further Work

9.1 Testing

Further corpus investigation is planned to determine likely coverage. In addition, experiments are planned using a text-based chat tool, to investigate both natural clarification question forms and responses thereto.

The corpus investigation is also intended to determine whether the requirement for syntactic parallelism of answers is acceptable.

9.2 Extensions

This approach can be naturally extended to deal with related underspecified phenomena, including unknown referents, lexical ambiguity and unknown senses of known words.

9.2.1 Unknown Referents

Unknown *referents* (of e.g. proper names, definite descriptions and anaphors), where it is the value of the INDEX feature of the parameter that is unknown (rather than the RESTR feature), can be treated by this approach with no modification except that new lexical entries for the resolved sign are not required.

9.2.2 Lexical Ambiguity

Lexically ambiguous words (those with more than one sense listed in the lexicon) should be treatable by this method by considering them

[18] Direct answers were defined as either directly parallel elliptical fragments, or answers of the form *"X means Y"*.

as “partially unknown” words: instead of the radically underspecified RESTR value as shown for an unknown word in AVM [4], a lesser degree of underspecification is required, as sketched out in AVM [10]:

$$[10]\quad \begin{bmatrix} \text{PHON} & \langle \text{bank} \rangle \\ \text{CONTENT} & \boxed{1} \\ \text{C-PARAMS} & \left\{ \boxed{1} \begin{bmatrix} \textit{parameter} & \\ \text{INDEX} & \boxed{i} \\ \text{RESTR} & \left\{ \begin{array}{l} \begin{bmatrix} \textit{noun-rel} & \\ \text{LEXSEM} & \text{bank}_{river} \\ \text{INSTANCE} & \boxed{i} \end{bmatrix} \vee \\ \begin{bmatrix} \textit{noun-rel} & \\ \text{LEXSEM} & \text{bank}_{finance} \\ \text{INSTANCE} & \boxed{i} \end{bmatrix} \end{array} \right\} \end{bmatrix} \right\} \end{bmatrix}$$

This underspecification could then be resolved in the same way as described above (or, if possible, resolved using dialogue context during the grounding process).

The use of logical disjunction (as shown in AVM [10]) to represent this underspecification may not be satisfactory (see e.g. (van Deemter, 1996) for a discussion of its inadequacy in standard approaches to ambiguity representation). However, it may be acceptable if a semantic representation that includes a level of illocutionary force is used (see e.g. (Ginzburg et al., 2001b)).[19] Further analysis must be performed, but the general approach should still be applicable if meta-level disjunction is required.

Whether the underspecification should be at the *parameter* level as shown here, or elsewhere (say, within the LEXSEM feature), remains to be determined.

9.2.3 New Senses of Old Words

As noted by (Garrod and Pickering, 2001), conversational participants may use words in innovative senses, but are able to align their understanding of the intended sense. If the grounding process can be designed to fail when the normal lexicon-specified senses cannot be anchored in context, the same clarification process could then be used to establish the intended sense.

One possible approach might be to allow creation of a new generic “unknown” lexical entry for a known word, once clarification has established the intended meaning. Another might be to consider known words as lexically ambiguous between their known sense(s) and an unknown sense which can be resolved as above.

9.3 Related Work

It is hoped that this approach can also be generalized to take in other phenomena that require clarification, e.g. unparseable utterances, presupposition failure and ambiguity of phonological identity (as often produced by speech recognisers).

10 Conclusions

The method described in this paper allows unknown words to be parsed (assigning underspecified lexical semantics) and subsequently acquired via a clarification question and answer exchange.

All operations are integrated within a HPSG grammar (together with dialogue update rules required to form the clarification question and new lexical entry), thus allowing this process to form part of a general dialogue strategy.

The method can be extended to apply to unknown referents, and seems to be applicable to known but ambiguous words.

11 Acknowledgements

The author is supported by the Engineering and Physical Sciences Research Council of the UK (grant number R000222969). He also wishes to thank Jonathan Ginzburg, Shalom Lappin, Aline Villavicencio and the ACL/SIGdial reviewers for many useful comments.

[19] Logical disjunction does not accurately represent ambiguity in a standard semantic representation – asserting *“I’m going to the bank”* is not the same as asserting *“I’m going to the riverside or the financial institution”*, so cannot be represented as:

$$go(X) \wedge (bank_r(X) \vee bank_f(X))$$

However, if the C-PARAMS are seen as being outside the scope of an illocutionary operator, the representation becomes:

$$\lambda\{X\}[bank_r(X) \vee bank_f(X)]assert(A, B, \langle\langle go(X)\rangle\rangle)$$

i.e. an assertion *“I’m going to the X”*, where the addressee knows that X is a riverside or a financial institution.

References

Petra Barg and Markus Walther. 1998. Processing unknown words in HPSG. In *Proceedings of COLING-ACL'98*, volume 1, pages 91–95.

Lou Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.

David Carter. 1992. Lexical acquisition. In H. Alshawi, editor, *The Core Language Engine*, pages 217–234. MIT Press, Cambridge, MA.

Ann Copestake, Dan Flickinger, Ivan Sag, and Carl Pollard. 1999. Minimal recursion semantics: An introduction. Draft.

Gregor Erbach. 1990. Syntactic processing of unknown words. In P. Jorrand and V. Sgurev, editors, *Artificial Intelligence IV – methodology, systems, applications*. North-Holland, Amsterdam.

Simon Garrod and Martin Pickering. 2001. Toward a mechanistic psychology of dialogue: The interactive alignment model. In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *Proceedings of the Fifth Workshop on Formal Semantics and Pragmatics of Dialogue*. BI-DIALOG.

Jonathan Ginzburg and Robin Cooper. 2001. Resolving ellipsis in clarification. In *ACL/EACL01 Conference Proceedings*. Association for Computational Linguistics, July.

Jonathan Ginzburg and Robin Cooper. forthcoming. Clarification, ellipsis and utterance representation.

Jonathan Ginzburg, Howard Gregory, and Shalom Lappin. 2001a. SHARDS: Fragment resolution in dialogue. In H. Bunt, I. van der Sluis, and E. Thijsse, editors, *Proceedings of the Fourth International Workshop on Computational Semantics (IWCS-4)*, pages 156–172. ITK, Tilburg University, Tilburg.

Jonathan Ginzburg, Ivan A. Sag, and Matthew Purver. 2001b. Integrating conversational move types in the grammar of conversation. In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *Proceedings of the Fifth Workshop on Formal Semantics and Pragmatics of Dialogue (BI-DIALOG 2001)*, pages 45–56.

Richard H. Granger. 1977. FOUL-UP: A program that figures out meanings of words from context. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI-77)*, volume 1, pages 172–178, August.

Howard Gregory. 2001. A ProFIT grammar and resolution procedure for fragments in dialogue. Technical Report TR-01-03, Department of Computer Science, King's College London, May.

Peter Hastings. 1994. *Automatic Acquisition of Word Meaning from Context*. Ph.D. thesis, University of Michigan.

Kevin Knight. 1996. Learning word meanings by instruction. In *Proceedings of the Thirteenth National Conference on Artifical Intelligence*, pages 447–454. AAAI/IAAI.

Ted Pedersen. 1995. Automatic acquisition of noun and verb meanings. Technical Report 95-CSE-10, Southern Methodist University, June.

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

James Pustejovsky. 1998. The semantics of lexical underspecification. *Folia Linguistica*, 32(3–4):323–347.

Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*. University of Pennsylvania, May.

Dale W. Russell. 1993. *Language Acquisition in a Unification-Based Grammar Processing System using a Real-World Knowledge Base*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Cynthia Thompson. 1998. *Semantic Lexicon Acqusition for Learning Natural Language Interfaces*. Ph.D. thesis, University of Texas at Austin, December.

David Traum, Johan Bos, Robin Cooper, Staffan Larsson, Ian Lewin, Colin Matheson, and Massimo Poesio. 1999. A model of dialogue moves and information state revision. In *Task Oriented Instructional Dialogue (TRINDI): Deliverable 2.1*. University of Gothenburg.

Kees van Deemter. 1996. Towards a logic of ambiguous expressions. In K. van Deemter and S. Peters, editors, *Semantic Ambiguity and Underspecification*, number 55 in CSLI Lecture Notes. CSLI Publications.

Uri Zernik. 1987. Language acquisition: Learning a hierarchy of phrases. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87)*, volume 1, pages 125–132, August.