An Empirical Study of Errors in Translating Natural Language into Logic

Dave Barker-Plummer (dbp@stanford.edu)
CSLI, Stanford University
Stanford, California, 94305, USA

Richard Cox (richc@sussex.ac.uk)
Department of Informatics, University of Sussex
Falmer, E. Sussex, BN1 9QJ, UK

Robert Dale (rdale@ics.mq.edu.au)
Centre for Language Technology, Macquarie University
Sydney, NSW 2109, Australia

John Etchemendy (etch@csli.stanford.edu)
CSLI and Philosophy, Stanford University
Stanford, California, 94305, USA

Abstract

Every teacher of logic knows that the ease with which a student can translate a natural language sentence into formal logic depends, amongst other things, on just how that natural language sentence is phrased. This paper reports findings from a pilot study of a large scale corpus in the area of formal logic education, where we used a very large dataset to provide empirical evidence for specific characteristics of natural language problem statements that frequently lead to students making mistakes. We developed a rich taxonomy of the types of errors that students make, and implemented tools for automatically classifying student errors into these categories. In this paper, we focus on three specific phenomena that were prevalent in our data: Students were found (a) to have particular difficulties with distinguishing the conditional from the biconditional, (b) to be sensitive to word-order effects during translation, and (c) to be sensitive to factors associated with the naming of constants. We conclude by considering the implications of this kind of large-scale empirical study for improving an automated assessment system specifically, and logic teaching more generally.

Keywords: errors; slips; misconceptions; natural language; e-learning; human reasoning; automated assessment; educational data mining; first-order logic; propositional logic; Language, Proof & Logic

Introduction

It seems obvious that the difficulty students face in translating natural language statements into formal logic will, at least in part, be due to characteristics of the natural language statements themselves. For example, we would expect it to be relatively easy to translate a natural language sentence when the mapping from natural language into logical connectives is transparent, as in the case of the mapping from *and* to '$\wedge$', but harder when the natural language surface form is markedly different from the corresponding logical form, as in the translation of sentences of the form *A provided that B*. However, evidence for this hypothesis is essentially anecdotal, and we have no quantitative evidence of *which* linguistic phenomena are more problematic than others.

This paper presents results from a pilot study using a large-scale corpus in the area of first-order logic teaching at the undergraduate level. The corpus consists of student-generated solutions to exercises in *Language, Proof and Logic* (LPL; Barwise, Etchemendy, Allwein, Barker-Plummer & Liu, 1999), a courseware package consisting of a textbook together with desktop applications which students use to complete exercises.[1] Students may submit answers to 489 of LPL's 748 exercises[2] to The Grade Grinder (GG), a robust automated assessment system that has assessed approximately 1.8 million submissions of work by more than 38,000 individual students over the past eight years; this population is drawn from approximately a hundred institutions in more than a dozen countries. These submissions form an extremely large corpus of high ecological validity which we wish to exploit in order (*inter alia*) to gain insights into cognitive processes during formal reasoning (such as those associated with natural-to-formal language translation and interpretation), to extend our research on individual differences in reasoning (e.g. Stenning & Cox, 2006), to improve logic teaching, and, eventually, to enrich the Grade Grinder's feedback to students. Understanding the nature of students' errors is central to these aims; the corpus offers considerable scope for analyses with this in mind.

As a pilot study in how this rich data set might be used to inform logic teaching, we selected a single exercise involving the translation of twenty sentences from English into propositional logic. We developed and validated a taxonomy for categorising students' translation errors and used it as the basis for an automated error-classification system. We analysed several thousand solutions to the exercise and singled out three high-frequency types of error — antecedent–consequent reversals, substitutions of connectives, and substitutions of constants — for in-depth followup. All three demonstrate how particular aspects of the form of a natural language sentence impact on the ease with which students can translate this sentence into logic. Below, we describe this analysis in detail, and provide some conclusions regarding the implications of this work for improving an automated assessment system specifically, and logic teaching more generally.

Sampling

For the purposes of initial exploration we selected a natural language (NL) to first-order logic (FOL) translation exercise of moderate difficulty, i.e. one that psychometrically discriminates between students. Exercise 7.12 from Chapter 7 (which introduces conditionals) was selected by computing the number of GG submissions per LPL exercise and rank ordering them by the proportion of incorrect submissions of the exercise. This exercise involves translating each of twenty English sentences into propositional logic (a subset of FOL). A

[1] See `http://lpl.stanford.edu`.

[2] The other exercises require that students submit their answers on paper to their instructors.

Translate the following English sentences into FOL. Your translations will use all of the propositional connectives.

1. *If* **a** *is a tetrahedron then it is in front of* **d**.
2. **a** *is to the left of or right of* **d** *only if it's a cube.*
3. **c** *is between either* **a** *and* **e** *or* **a** *and* **d**.
4. **c** *is to the right of* **a**, *provided it (i.e.* **c**) *is small.*
5. **c** *is to the right of* **d** *only if* **b** *is to the right of* **c** *and left of* **e**.
7. *If* **b** *is a dodecahedron, then it's to the right of* **b** *if and only if it is also in front of* **b**.
10. *At least one of* **a**, **c**, *and* **e** *is a cube.*
11. **a** *is a tetrahedron only if it is in front of* **b**.

Figure 1: An extract from Exercise 7.12

translation for a sentence (which we refer to here as a **solution**) is considered correct if it is equivalent to a **reference solution**.[3] Sample sentences from Exercise 7.12 are presented in Figure **??**.

The reference solution for Sentence 1 in Figure **??** is $\mathsf{Tet}(\mathsf{a}) \rightarrow \mathsf{FrontOf}(\mathsf{a},\mathsf{d})$. The Grade Grinder's response to an erroneous submission of the form $\mathsf{FrontOf}(\mathsf{a},\mathsf{d}) \rightarrow \mathsf{Tet}(\mathsf{a})$, a common error, takes the form:

```
*** Your first sentence, "FrontOf(a, d) ->
Tet(a)", is not equivalent to any of the
expected translations.
```

Further information on the Grade Grinder, and samples of feedback reports, can be found on the GG website.[4]

A **submission** for Exercise 7.12 consists of a solution for all twenty sentences, and is considered erroneous if the student makes an error on at least one of the solutions. We examined the corpus of submissions of Exercise 7.12 made by students during the calendar years 2000–2007 — more than 74,000 submissions, of which 42,416 submissions (57%) were erroneous, with a total of 148,681 incorrect translation solutions.[5] These submissions were made by 11,925 different students, representing an average of 12.47 incorrect translations per student.

Method: Developing an Error Taxonomy

We scrutinised a subsample of 296 erroneous solutions for Sentence 1 in Exercise 7.12, and worked collaboratively to develop a coding scheme for annotating students' solutions.

The errors in a student's answer are determined by comparison with the reference solution. Throughout this paper we assume that this answer is the most natural solution, and that it is this sentence that the student is aiming to produce.

The sentences $\mathsf{Tet}(\mathsf{a}) \rightarrow \mathsf{FrontOf}(\mathsf{a},\mathsf{d})$ and $\neg\mathsf{Tet}(\mathsf{a}) \vee \mathsf{FrontOf}(\mathsf{a},\mathsf{d})$ are equivalent, correct translations for Sentence 1; here, the first of these is the reference solution. A student submission of $\mathsf{Tet}(\mathsf{a}) \vee \mathsf{FrontOf}(\mathsf{a},\mathsf{b})$ will be classified as a connective substitution of '$\vee$' for '$\rightarrow$', while it is equally plausible that the student was aiming at the second answer, and committed the error of omitting the '$\neg$'.

A framework, or set of guidelines, emerged following an iterative, fairly systematic process based on the procedure advocated by Chi (1997) for developing protocols for the qualitative analysis of verbal data in educational research. We identified, grouped and categorised broad classes of error types, iteratively developed a coding scheme, and wrote operational definitions of the error types. These processes were cyclical and repeated at various levels of granularity. Eventually, a taxonomy of 45 error types emerged. Each of these is identifiable in terms of surface-form characteristics: we also explored the categorisation of the different errors in terms of their possible *causes*, but decided that this left too much scope for subjectivity, and that a categorisation based on observable phenomena was most appropriate at this stage in the exercise. We organised these 45 error types under three broad categories, based on the representation of a FOL formula as a tree:

1. **Structural Errors**: these are errors involving the structure of the FOL tree—for example, switching of the antecedent and consequent of an implication, or adding exclusivity (e.g. giving a sentence of the form $(\mathsf{A} \vee \mathsf{B}) \wedge \neg(\mathsf{A} \wedge \mathsf{B})$ instead of just $(\mathsf{A} \vee \mathsf{B})$.

2. **Connective Errors**: errors involving labels on the interior nodes of the FOL tree: using one connective in place of another.

3. **Atomic Errors**: errors involving the structure of an atomic subformula; these are of two subtypes:

   (a) **Predicate Errors**, where one predicate symbol is used in place of another; and

   (b) **Argument Errors**, where one argument is used in place of another, or the wrong number of arguments is present.

Each of the categories cover a collection of different kinds of error. For example, **Connective Errors** can be the substitution of a connective for any other ('$\wedge$' in place of '$\rightarrow$', say), for an out-of-vocabulary (unknown) symbol, or for the empty symbol (omission). For **Argument Errors**, we distinguish the situation where a constant has been substituted for the correct symbol in the wrong orthographic case ('A' instead of 'a', say) from substitutions of one constant for another.

Additional error types derive from stereotypical patterns of lower-level errors. For example, a substitution of one constant for another throughout the formula is categorised as a 'uniform' substitution, while substitution in some places and not others is categorised as 'sporadic'. We also introduce a 'waste-basket' class covering cases where students have introduced unnecessary or unmatching parentheses; inclusion of periods at the end of strings; and examples so radically different from the solution that it is not clear how to characterise them.

[3] There are infinitely many correct answers for any sentence, so a theorem prover is employed to determine equivalence.

[4] See http://ggww2.stanford.edu/GUS/lpl/.

[5] Some or all of the translations may be empty and therefore erroneous, but we ignore these null solutions in this paper.

Table 1: Examples of errors, (*n.b.* ACREV = Antecedent-Consequent Reversal)

#	Reference solution	Errored solution	Type	Subtype
1	$\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,d)}$	$\mathsf{FrontOf(a,d)} \rightarrow \mathsf{Tet(a)}$	1	ACREV
2	$\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,d)}$	$\mathsf{FrontOf(a,b)} \rightarrow \mathsf{Tet(a)}$	1, 3ii	ACREV, Incorrect Constant
3	$\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,d)}$	$\mathsf{Tet(a)} \vee \mathsf{FrontOf(a,d)}$	2	Disjunction for Conditional
4	$\neg\mathsf{Cube(e)} \rightarrow (\mathsf{Large(b)} \vee \mathsf{Large(d)})$	$\neg\mathsf{Cube(e)} \rightarrow \mathsf{Large(b)} \vee \mathsf{Large(d)}$	1	Missing Parens
5	$\mathsf{Large(e)} \rightarrow \mathsf{Large(a)}$	$\mathsf{e} \rightarrow \mathsf{Large(a)}$	2	Elided Predicate
6	$\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,d)}$	$\mathsf{Tet(a)} \rightarrow \mathsf{InFrontOf(a,d)}$	3i	Incorrect Predicate
7	$\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,d)}$	$\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,b)}$	3ii	Incorrect Constant
8	$\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,d)}$	$\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(d)}$	3ii	Arity Error

Table **??** shows examples of each of the higher-level error types, each labelled with the more specific error category assigned in our taxonomy. Note that some solutions, as in example 2 here, contain more than one error.

Reliability of the Coding Scheme

For manageability in annotation, we identified and characterised eight error types that serve as intermediate nodes in the error taxonomy between the three top level types and the 45 leaf-node categories. Two independent annotators used this categorisation to code 296 student solutions to Sentence 1 of Exercise 7.12. Each solution contained between one and three errors. We computed Cohen's kappa (κ) statistic for inter-annotator reliability; kappa can be used to verify that agreement exceeds chance levels (e.g. Viera & Garrett, 2005). κ was computed separately for each of the categories; values are shown in parentheses after each category: Antecedent–Consequent Reversal (κ=.97); Incorrect Constants (κ=.98); Incorrect Predicates (κ=.74); Incorrect Connectives (κ=.94); Parenthesis Errors (κ=.76); Case Errors (κ=.66); Arity Errors (κ=1.00); Other Errors (κ=.52). *Kappa* values all showed substantial, almost perfect or perfect agreement for all categories except 'Other', for which there was only moderate agreement.

Automating the Coding Process

We used the full taxonomy to inform the design of an automated solution coding system. We developed a simple pattern-matching program, based on regular expressions, and used it to classify automatically each of the sentences in the complete corpus. The classifier identifies fragments of the submitted answer that appear to correspond to the atomic subformulae of the reference sentence, and requires that the correct number of such subformulae are present before proceeding to classify the errors within the answer. This approach allows us to analyse those answers which are syntactically incorrect, although it does not have the flexibility that a tokenizer or chart-parser based analyser might have.

On average the classifier was able to classify 85% of the submissions within the corpus. The classification rate varied with sentence and ranged from 62% (Sentence 7) to 99% (Sentence 11). The classifier was able to code for each of the errors in the taxonomy, but the presence of some errors prevents the coding for some others. For example, if a submission contains the wrong predicate (Incorrect Predicate error), we do not subsequently check that the arity of that predicate is correct (Arity Error).

Table 2: Error frequencies

Error Type	Count	%age of All
Antecedent–Consequent Reversal	25084	25.86%
Biconditional for Conditional	17518	18.06%
Conditional for Biconditional	11362	11.71%
Negation Error	8954	9.23%
Incorrect Scope	5422	5.59%
Failure to Scope	4701	4.85%
Argument Error	4474	4.61%
Conjunction for Conditional	3187	3.29%
Conditional for Conjunction	2091	2.16%
Biconditional for Conjunction	1514	1.56%

Applying the Taxonomy to Data

We ran the student's FOL translation solutions to the twenty Exercise 7.12 sentences through the regular expression-based coding software. We first produced a list of frequencies of errors for all 20 sentences, and focussed our attention on the 10 most frequent errors for each sentence. The results of this analysis are shown in condensed form in Table **??**, which shows the number of instances of each of the ten most frequent error types across all 20 sentences, along with the percentage of total errors accounted for by these instances.

Below, we illustrate the application of the taxonomy to the data with one example from each of the three broad categories mentioned earlier (Structural Errors, Connective Errors and Atomic Errors): Antecedent–Consequent Reversal and Incorrect Substitution of the Biconditional for the Conditional are the two most frequent errors overall, and Argument Errors are the most common Atomic Errors.

Reversal of Antecedent and Consequent

Clement, Lochhead and Monk (1981) studied translation difficulties in mathematics. Students talked aloud as they worked on a simple algebra problem ("Write an equation: 'There are six times as many students as professors at this university.' Use **S** for the number of students and **P** for the number of professors".). The predominant error consisted of re-

versing the variables in the equation. Of 150 'calculus level' students, 37% manifested reversal errors of the form $6S = P$. The rate was 57% in another sample. Two sources for the error were identified. The first was termed 'word-order matching' in which the student orders terms in their equation in a way that matches the order of keywords in the problem statement. The process is superficial and syntactic. Another strategy Clement *et al*. (1981) termed 'static comparison'. This is one in which the student does not understand **S** as a variable representing the number of students, but as a label attached to a number, in this case '6'. The student places a multiplier adjacent to the letter associated with the larger group. This error is based on an interpretation of the equals sign as expressing comparison or association rather than equality. Clement (1982) reports that the static comparison strategy is a 'deep-seated, intuitive symbolization strategy ...' which can co-exist with formally taught contradictory schemes and which can 'take over' in some problem solving situations (p.28). We were interested to investigate whether evidence for strategies akin to word-order matching and static comparison in word algebra problem solving can also be found in the FOL domain.

There are two kinds of sentence for which element ordering matters (i.e. where the connective is the non-commutative '$\rightarrow$'). For some, the correct solution preserves the word ordering in the posed NL sentence (e.g. Sentence 1: *If a is a tetrahedron then it is in front of d* $\approx \mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,d)}$). For others, the correct solution requires re-ordering (e.g. Sentence 4: *c is to the right of a, provided it (i.e., c) is small* $\approx \mathsf{Small(c)} \rightarrow \mathsf{RightOf(c,a)}$). Twelve of our reference sentences involve implication; eight of them preserve word order and four do not. Our prediction was that Antecedent–Consequent Reversal would occur more frequently on sentences for which the correct solution requires re-ordering than would be observed for sentences that preserve the posed word order in the solution. We calculated the number of erroneous solutions submitted to Grade Grinder which demonstrated Antecedent–Consequent Reversal for each sentence and expressed that as a fraction of the total number of solutions for that sentence. The total number of solutions submitted across the 12 sentences in the analysis ranged from 318 to 8361.

For the four sentences in which word order is not preserved during translation from posed NL sentence to FOL solution, the percentage of antecedent and consequent reversals was 66%. This means that one basis for the students' error was a tendency to preserve the original word order for this type of sentence. In contrast, for the eight sentences in which word order is preserved during translation, the antecedent and consequent reversal rate was 43%. This difference was significant under a directional hypothesis (i.e. that sentences requiring re-ordering during translation would produce more Antecedent–Consequent Reversals).

The t-test values were ($t = 2.23$, $df(10)$, $p = .05$, 2-tailed). It therefore seems that word order in the NL posed sentence does tend to 'drive' term order in the FOL translation for many students, though this effect is probably conflated with difficulties in distinguishing the biconditional from the conditional, as discussed in the next section.

Table 3: Biconditional for Conditional Errors

Frequency	Percentage	Surface Form
13214	75.43%	S only if S.
1777	10.14%	S unless S.
1146	6.54%	S provided S.
725	4.14%	S if S.
367	2.09%	If S then if S then S.
289	1.65%	If S then S.

In future studies we plan to look at correlations of error patterns within and between students; this may help to establish the relative contribution of each phenomenon. Consequently, they provide an interesting means by which the effect of natural language presentation upon connective substitution errors can be investigated.

Connective Substitutions

Sentences 1 and 11 have quite different NL forms, but their FOL translations are identical modulo the use of a different constant.

1. *If a is a tetrahedron then it is in front of d*, $\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,d)}$;
11. *a is a tetrahedron only if it is in front of b*, $\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,b)}$

Of the 10 most common errors for these two sentences, six are shared by both: Antecedent–Consequent Reversal, Biconditional for Conditional substitution, Argument Reversal, Incorrect Argument, Incorrect Predicate, and Arity Error. The frequency of occurrence of each of the six shared error types were rank-ordered 1–6 separately for each of Sentences 1 and 11.

Whereas for Sentence 1 Biconditional for Conditional substitution ranked fourth out of the six and represented 12% of 1361 solutions, for Sentence 11 it ranked first — the most common error of all, with 58% of 8981 solutions of this kind. This strongly indicates that there is something about the surface structure of the two NL sentences that elicit very different error patterns from students. More generally, we can consider the propensity for students to make the Biconditional for Conditional substitution error when faced with a variety of different natural language renderings of the conditional. Table **??** shows the number and percentage of Biconditional for Conditional errors per surface form across the 20 sentences. This demonstrates that students find it significantly more difficult to translate the *only if* form than other natural renderings of the conditional.

Substitution of Constants

Incorrect Constant errors, where one constant is substituted for another, are a significant form of error for specific sentences. For example, four out of the top 10 error forms for Sentence 10 (*At least one of* a, c, *and* e *is a cube* $\approx \mathsf{Cube(a)} \vee \mathsf{Cube(c)} \vee \mathsf{Cube(e)}$) involve this type of error:

$\mathsf{Cube(a)} \vee \mathsf{Cube(b)} \vee \mathsf{Cube(c)}$	$n = 758$
$\mathsf{Cube(a)} \vee \mathsf{Cube(b)} \vee \mathsf{Cube(e)}$	$n = 227$
$(\mathsf{Cube(a)} \vee \mathsf{Cube(b)} \vee \mathsf{Cube(c)})$	$n = 53$
$\mathsf{Cube(a)} \vee \mathsf{Cube(c)} \vee \mathsf{Cube(b)}$	$n = 45$

We noted that this kind of error seemed to interact with (1) the use of the constant a in a sentence; (2) whether the use of a as a constant was the first mentioned constant in the sentence; and (3) whether or not the letters used as constant names were alphabetically adjacent (e.g. ⟨a, b, c⟩) versus whether the sentence's constant letters were alphabetically 'gappy' (e.g. ⟨b, e, d⟩). Visual inspection of the data suggested at least two trends: the first was for constant substitutions to be more frequent when the letters used as constants were not alphabetically adjacent, and the second was for this effect to appear to be magnified when the letters used as constants were not alphabetically adjacent and the first constant letter name mentioned in the sentence was a.

To investigate these issues, we binary-coded each of the sentences in terms of these factors (present/absent). These were used as independent variables in analyses with the constant substitution frequency data as the dependent variable. An independent t-test comparing the normalised mean frequencies of constant substitutions for 'gappy' versus 'non-gappy' sentences revealed a highly significant difference ($t = 3.58$, $df(13.4)$, $p < .005$). Non-gappy sentences ($n = 6$) averaged 4% constant substitution errors, whereas 20% of the errors on gappy sentences ($n = 14$) were constant substitution errors, usually of the ⟨a, b⟩ for ⟨a, d⟩ variety. On Sentence 1, for example (see Figure **??**), 85% of constant substitutions were of b for d, compared to only 25% for d for b substitutions on Sentence 11.

We also compared sentences in which the constant name a was used with those in which it was not, and whether or not it was mentioned as the first constant in the NL sentence. The effect of this factor was also significant under a 1-tailed t-test ($t = 2.00$, $df(8)$, $p < .05$). Twenty-four percent of errors on sentences with a mentioned as first constant name were constant substitution errors, compared to only 9% for other sentences. The interaction of the gappy and 'a-first' factors approached statistical significance and suggested a tendency for the gap effect to be magnified in sentences in which a is the first mentioned constant name (Table **??**).

Discussion

The results of the analyses presented here provide support for the hypothesis that properties of the surface form of a natural language sentence negatively impact translation performance when the surface form differs markedly from the corresponding logical form. As we have demonstrated, automated analysis of a very large data set allows the exploration of specific hypotheses regarding the effects of particular aspects of surface form. We have shown that surface features such the ordering of antecedent and consequent terms (discussed above), the use of particular connectives, and the way in which constants are named affect translation performance. In this section we discuss the latter two features in more detail.

Table 4: Constant substitution errors as %s of all errors for sentences in which first mentioned constant was/was not a vs. whether or not constant names were alphabetically adjacent letters ('gappyness').

Begins with 'a'?	Gappy?	N	Mean constant substitution error (%)
no	no	4	4.5%
no	yes	8	11%
yes	no	2	2%
yes	yes	6	32%

Connective Substitution

It seems plausible to suggest that the use of *only if* in Sentence 11 cues the phrase *if and only if* in the student's mind. The meaning of the term is introduced in the LPL textbook (p.180) as follows: "... the expression *only if* introduces ... a *necessary condition*". The text provides an illustration involving an instructor saying to the class that "you will pass the course only if you turn in all the homework assignments", pointing out that this does *not* imply that if all homework is handed in passing is guaranteed. The biconditional is introduced and illustrated in a similar vein. One source of the difficulty that students appear to have with distinguishing the conditional and the biconditional may stem from the way in which *if* is used in natural language. For example, Stenning and van Lambalgen (2001), citing Geis and Zwicky (1971), argue that conditionals are often naturally interpreted as biconditionals in everyday contexts especially where conditions are implied or ultimatums are issued (deontics). They suggest that a statement such as "if you read this, I'll buy you lunch" 'drops a heavy hint that no reading, no lunch' (p. 287). Stenning and van Lambalgen (2001) also point out that this kind of interpretation is akin to the Gricean maxim of *relevance* (Grice, 1975) under which, if the hearer assumes that if his interlocutor was going to buy lunch anyway, then why would she make the promise conditional upon the performance of some task?

Constant Substitution as a Capture Error

We hypothesise that constant substitution is sometimes a **slip**[6] of the 'capture error' type, in which a more frequent behaviour 'captures' a less frequent behaviour. For example, we sometimes might dial a frequently-dialled number when we intended to dial a number beginning with the same prefix. Under our hypothesis, the more frequent behaviour is the use of alphabetical names in order: ⟨a, b, c, ...⟩, and that this behaviour is capturing the required usage: ⟨a, c, e⟩ for Sentence 10. The data presented in table **??** support this hypothesis. The presence of a as the initial constant appears to prime the

[6] A term used in the human error literature. e.g. Reason (1990)

familiar behaviour, which results in a high level of constant substitution when 'gaps' too are present.

Future Work

Our initial explorations of this large data set have produced a plethora of interesting directions to pursue.

In the foregoing, we have explored the correlation between specific surface forms and specific error types. There are more complex language-related aspects worthy of exploration. For example, we might characterise the NL sentences in terms of their 'naturalness' or 'paraphrase distance' from their FOL translated form. Sentence 12 (**b** *is larger than both* **a** *and* **e**) seems quite 'everyday' in its phrasing compared to sentence 10 (*at least one of* **a**, **c** *and* **e** *is a cube*), for example.

We can also consider the complexity of the natural language forms in terms of factors such as: the number of clauses they contain; whether or not these clauses are embedded; whether there is scope for ambiguity in pronominal reference resolution or the resolution of elided elements; whether the sentences contain conjunction, negation, or other signals of syntactic complexity. A better understanding of the impact of these factors on the difficulty of translation could lead, for example, to the automatic generation of natural language sentences that test a student's specific weaknesses. We also seek to clarify what cognitive processes give rise to the types of errors that we observe. The translation findings we report represent comprehension processes rather than full deductive inferential processes. We feel that the theoretical implications of our findings for abstract rule versus mental model theories must await analyses of 'deeper' deductive inferential reasoning (*e.g.* across several sentences), a focus of our current work.

Another aim is to investigate what interventions are most appropriate for instances of the different kinds of errors within the taxonomy. Instead of responding to all incorrect answers with a bald statement of fact as currently delivered by the Grade Grinder, we might instead report on the presence of substitution errors in a different way to antecedant–consequent reversal, for example.

In addition to cross-subject analyses such as those presented here, the Grade Grinder error corpus affords us the opportunity to examine within-subject effects, since we can identify sequences of attempts by a single student to solve an exercise, or sequence of exercises. As an intervention, we may be able to present a student with a profile of the errors that they are prone to commit, together with appropriate advice for avoiding those errors.

Conclusion

Our results have illustrated how it is possible to use empirical data gathered on a large scale to gain insights into the difficulty that students have when learning to translate sentences from FOL into NL. We believe that these insights can improve the standard of logic teaching (with or without the use of LPL, or software support) and other related areas such as mathematics and computer science. General observations such as those concerning the cooperative stance of natural language vs. the more adversarial stance required in these more formal fields will have wide application. If educators are trained to expect and recognize errors stemming from these causes, and have appropriate interventions available, then the quality of education in these areas may be improved. The potential value of greater understanding here is not limited to the teaching of logic - it is also an important component of 'computational thinking' (Wing, 2006). She defines it as 'a universally applicable attitude and skill set that everyone ... would be eager to learn and use" (p33); it includes, among many others, skills in problem decomposition and heuristic reasoning. An understanding of logic can facilitate these skills—abilities which are becoming ever more important if individuals are to benefit from technological developments in the modern world.

Acknowledgements

RC gratefully acknowledges the support provided by an ESRC Teaching & Learning Research Programme (UK) and SSRC (US) Visiting Americas Fellowship; RD gratefully acknowledges the support of the Australian Research Council. Albert Liu assisted with the collection of the data; Michael Murray and Dawit Meles helped with this pilot study.

References

Barwise, J., Etchemendy, J., Allwein, G., Barker-Plummer, D. & Liu, A. (1999) *Language, Proof and Logic.* CSLI Publications and University of Chicago Press.

Clement, J. (1982) Algebra word problems: Thought processes underlying a common misconception. *Journal for Research in Mathematics Education*, **13(1)**, 16–30.

Clement, J., Lochhead, J., & Monk, G.S. (1981) Translation difficulties in learning mathematics. *The American Mathematical Monthly*, **88(4)**, 286–290.

Chi, M. T. H. (1997) Quantifying qualitative analyses of verbal data. *Journal of the Learning Sciences*, **6(3)**, 271–315.

Geis, M.C. & Zwicky, A.M. (1971) On invited inferences. *Linguistic Enquiry*, **2**, 561-566.

Grice, H.P. (1975) Logic and conversation. In P. Cole & J. Morgan, (eds.), *Syntax and Semantics: Vol 3. Speech Acts*. London: Academic Press.

Reason, J. (1990) Human Error. Cambridge, UK: Cambridge University Press.

Stenning, K. & Cox, R. (2006) Reconnecting interpretation to reasoning through individual differences. *The Quarterly Journal of Experimental Psychology*, **59 (8)**, 1454–1483.

Stenning, K. & van Lambalgen, M. (2001) Semantics as a foundation for psychology: A case study of Wason's selection task. *Journal of Logic, Language and Information*, **10**, 273–317.

Viera, A.J. & Garrett, J.M. (2005) Understanding interobserver agreement: The kappa statistic. *Family Medicine*, **37(5)**, 360–363.

Wing, J. (2006) Computational thinking. *Communications of the ACM*, **49(3)**, 33-35.