Student Translations of Natural Language into Logic: The Grade Grinder Corpus Release 1.0

Dave Barker-Plummer, Richard Cox and Robert Dale

Students find logic hard. In particular, they seem to find it hard to translate natural language sentences into their corresponding representations in logic. As an enabling step towards determining why this is the case, this paper presents the public release of a corpus of over 4.5 million translations of natural language (NL) sentences into first-order logic (FOL), provided by 55,000 students from almost 50 countries over a period of 10 years. The translations, provided by the students as FOL renderings of a collection of 275 NL sentences, were automatically graded by an online assessment tool, the Grade Grinder. More than 604,000 are in error, exemplifying a wide range of misunderstandings and confusions that students struggle with. The corpus thus provides a rich source of data for discovering how students learn logical concepts and for correlating error patterns with linguistic features. We describe the structure and content of the corpus in some detail, and discuss a range of potentially fruitful lines of enquiry. Our hope is that educational data mining of the corpus will lead to improved logic curricula and teaching practice.

1. INTRODUCTION

From a student's perspective, logic is generally considered a difficult subject. And yet it is an extremely valuable and important subject: the ability to reason logically underpins the Science, Technology, Engineering and Mathematics (STEM) fields which are seen as central in advanced societies. We believe it is in society's interests to make logic accessible to more students; but to do this, we need to have an understanding of precisely what it is about logic that is hard, and we need to develop techniques that make it easier for students to grasp the subject.

One key component skill in the understanding of logic is a facility for manipulating formal symbol systems. But such a skill is abstract and of little value if one does not also have the ability to translate everyday descriptions into formal representations, so that the formal skills can be put to use in real-world situations. Unfortunately, translating from natural language into logic is an area where students often face problems.

It seems obvious that the difficulties students face in this translation task 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.

It is against this background that we present in this paper the release of a publicly-available anonymised corpus of more than 4.5 million translations of natural language (NL) sentences into first-order logic (FOL) sentences, of which more than 604,000 (approximately 13%) are categorized by an automatic assessment tool as being in error. For each item in the corpus, we know what NL sentence was being translated, and we have both the FOL translation the student provided, and a 'gold-standard' answer representing the class of correct

Author's addresses:
Dave Barker-Plummer, CSLI, Cordura Hall, Stanford University Stanford, CA, 94305, USA; email: dbp@stanford.edu;
Richard Cox, Department of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, UK; email: rcox@inf.ed.ac.uk;
Robert Dale, Center for Language Technology, Department of Computing, Macquarie University, Sydney, NSW, 2109, Australia; email: Robert.Dale@mq.edu.au.

answers.[1] Students are identified by unique anonymised IDs, so the corpus allows us to determine how many previous attempts the student has made at the same exercise and the time intervals between attempts, and also to correlate any given student's performance across exercises. The data thus makes possible a broad range of analyses of student behaviors and performance. We are making the corpus available to the wider community in the hope that this will encourage research that leads to improvements in the teaching of logic.[2]

Section 2 explains the wider context in which this data has been collected, which has allowed us to gather a very large corpus of data regarding student performance at various tasks in logic learning. Section 3 then describes the focus of this paper—what we call the *Translations Subcorpus*—in more detail. Section 4 describes the format of the data as it appears in the corpus. Section 5 provides summary statistics over the errors in the corpus, and makes some observations about the nature of these errors. Section 6 concludes with some illustrative analyses and suggestions for ways in which this corpus can be exploited.

2. BACKGROUND

The data described here consists of student-generated solutions to exercises in *Language, Proof and Logic* (LPL; [Barwise et al. 1999]), a courseware package consisting of a textbook together with desktop applications which students use to complete exercises.[3] The LPL textbook is divided into three parts covering, respectively, Propositional Logic, First-Order Logic and Advanced Topics. The first two parts cover material typical of introductory courses in logic. Students completing these two parts of the textbook will have been exposed to notions of syntax and semantics of first-order logic and a natural deduction–style system for constructing formal proofs. Each of these areas of the course are supported by a number of software applications which provide environments where students can explore the concepts being taught.

The LPL textbook contains 748 exercises, which fall into two categories: 269 exercises which require that students submit their answers on paper to their instructors, and 489 for which students may submit answers to the Grade Grinder, a robust online automated assessment system that has assessed approximately 2.75 million submitted exercises by more than 55,000 individual students in the period 2001–2010. This student population is drawn from approximately a hundred institutions in almost fifty countries. Figure 1 provides statistics on how this data breaks down across the 10 years that the corpus represents.[4]

Student users of the system interact with the Grade Grinder by constructing computer files that contain their answers to particular exercises that appear in the LPL textbook. These exercises are highly varied, and make use of the software applications packaged with the book. Some focus on the building of truth tables using an application called Boole; some involve building blocks world scenarios using a multimodal tool called Tarksi's World, in which the student can write FOL sentences and simultaneously build a graphical depiction which can be checked against the sentences; and some require the construction of formal proofs using an application called Fitch. The Grade Grinder provides us with significant collections of data in all these areas. The exercises of interest here are what we call **translation exercises**; they form the basis of the corpus whose release this paper describes, and we discuss them in detail in Section 3 below.

The Grade Grinder corpus is similar to some of the corpora in the PSLC Datashop repository [Koedinger et al. 2010]. It shares with these the characteristics of being extensive (millions of data points) and longitu-

[1] Since the same information can be expressed by many different FOL sentences, any answer that is provably equivalent to this gold-standard answer is considered correct.

[2] A website is under development; in the interim, the corpus may be obtained by contacting the authors. A longer version of this paper which describes the corpus in more detail is available as a technical report[Barker-Plummer et al. 2011].

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

[4] The 'Domains' column shows the number of different internet country domains found in the email addresses of the student population for the year in question; definitively correlating these with countries is difficult since a student may use an email address in a domain other than that of their home country, the international use of `.com` mail hosts being the most obvious instance.

Year	Submissions	Students	Instructors	Domains
2001	190,653	4,097	142	23
2002	237,942	5,219	152	26
2003	238,104	5,106	168	33
2004	251,898	5,473	196	28
2005	255,974	5,295	182	27
2006	266,208	5,295	207	31
2007	304,719	6,444	224	33
2008	322,273	7,174	243	31
2009	331,746	6,489	212	33
2010	352,262	7,404	217	23

Fig. 1. Grade Grinder Usage Statistics: 2001–2010

dinal (repeat submissions by students over a semester or longer). However, it is not as fine-grained as many DataShop datasets.[5] For example, the DataShop Geometry tutor dataset contains data on students' actions and system responses at the level of knowledge components (skills or concepts). In contrast, a Grade Grinder submission represents the end-point of a student's work on an exercise. The corpus described here also differs from many DataShop corpora in that is not derived from an intelligent tutoring system or cognitive tutor, but from a blended learning *package* consisting of courseware, several desktop computer applications, and an online grading system.

3. NATURAL LANGUAGE TO LOGIC TRANSLATIONS

As noted above, the exercises in LPL cover a range of different types of logic exercises, and so the Grade Grinder's collection of assessments is very large and varied. Over time, we aim to make the various components of this corpus available; as a first step, we are making available what we believe may be the most useful component of the corpus, this being the part that is concerned with students' translations of natural language sentences into logic.

Translation exercises ask the student to translate a number of what we will call **translatable sentences**, writing their answers in a single file, which is then submitted to the Grade Grinder. We will refer to each submission of a translated sentence as a **translation act**. Figure 2 shows an example exercise that calls for the student to translate twenty English sentences into the language of FOL. The student's response to such an exercise is considered correct if it contains a translation act for every translatable sentence in the exercise, and every translation act corresponds to a correct translation. The LPL textbook contains 33 translation exercises, involving a total of 275 distinct translatable NL sentences.

The Grade Grinder examines each submitted file, making a note of errors that are found within the student's answers. The files are saved to the corpus, the errors are noted, and an email message is sent to the submitter summarizing these errors. Currently, the Grade Grinder offers only *flag feedback* [Corbett and Anderson 1989], indicating only whether a submitted solution is correct. The software makes no attempt to diagnose the error that has been made, apart from reporting the difference between a well-formed expression of logic that is incorrect, and an ill-formed expression which is meaningless. Figures 3 and 4 respectively give examples of the feedback for the submission of correct and incorrect solutions to the exercise shown in Figure 2. The feedback report in Figure 4 indicates that the student has submitted an incorrect answer to the second sentence, and an ill-formed expression in answer to the sixth sentence. The solution for sentence eighteen is also reported as ill-formed, since there is no text in this slot of the solution.

Each student may submit solutions to the same exercise as many times as desired. Once a student is satisfied with their work, they may submit the work again, this time requesting that a copy of the system's

[5]However, note the File Timestamps information discussed in Section 4.

Exercise 7.12 (Translation) 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.
(6) *If* e *is a tetrahedron, then it's to the right of* b *if and only if it is also in front of* b.
(7) *If* b *is a dodecahedron, then if it isn't in front of* d *then it isn't in back of* d *either.*
(8) c *is in back of* a *but in front of* e.
(9) e *is in front of* d *unless it (i.e.,* e*) is a large tetrahedron.*
(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.
(12) b *is larger than both* a *and* e.
(13) a *and* e *are both larger than* c*, but neither is large.*
(14) d *is the same shape as* b *only if they are the same size.*
(15) a *is large if and only if it's a cube.*
(16) b *is a cube unless* c *is a tetrahedron.*
(17) *If* e *isn't a cube, either* b *or* d *is large.*
(18) b *or* d *is a cube if either* a *or* c *is a tetrahedron.*
(19) a *is large just in case* d *is small.*
(20) a *is large just in case* e *is.*

Fig. 2. An example exercise (7.12) from LPL

```
Grade report for Oedipa Maas (oedipa@yoyodyne-industries.com)
Submission ID: 11.076.18.28.21.L00-0002222
Submission received at: Thu Mar 17 18:28:21 GMT 2011
Submission graded at: Thu Mar 17 18:28:33 GMT 2011
Submission graded by: gradegrinder.stanford.edu

#### No instructor name was given.  The report was sent only to the student.

The following files were submitted:
        Sentences 7.12

EXERCISE 7.12

 Sentences 7.12  (Student file: "Sentences 7.12")
    Your sentences are all correct.  Hurrah!
```

Fig. 3. Example feedback from the Grade Grinder: A translation exercise without errors

email response be sent to a named instructor. The effect of this pattern of interaction with the Grade Grinder is that the corpus contains a trace of each student's progression from their initial submission to their final answer.

We can categorize the translation exercises along three dimensions as follows, and as summarized in Figure 5.

Logical Language. The LPL textbook introduces the language of first-order logic in stages, starting with atomic formulae in Chapter 1, then the Boolean connectives ($\wedge$, $\vee$ and $\neg$) in Chapter 3, followed by conditional connectives ($\rightarrow$ and $\leftrightarrow$) in Chapter 7. These connectives together define the propositional fragment of first-order logic. Finally, the universal and existential quantifiers ($\forall$,$\exists$) are introduced in Chapter 9 to complete the language of first-order logic. Exercises have correspondingly complex languages according to the position in which they appear.

```
Grade report for Tyrone Slothrop (tyrone@yoyodyne-industries.com)
Submission ID: 11.076.18.30.56.L00-0002222
Submission received at: Thu Mar 17 18:30:56 GMT 2011
Submission graded at: Thu Mar 17 18:31:02 GMT 2011
Submission graded by: gradegrinder.stanford.edu

#### No instructor name was given.  The report was sent only to the student.

The following files were submitted:
        Sentences 7.12

EXERCISE 7.12

 Sentences 7.12  (Student file: "Sentences 7.12")
    We found problems in your sentences:
    *** Your second answer, "~SameCol(a d)->Cube(a)", isn't well formed.
    *** Your sixth sentence, "Tet(e)->(RightOf(e, b)->FrontOf(e, b))", is not
        equivalent to any of the expected translations.
    *** Your fifteenth sentence, "Large(a)->Cube(a)", is not equivalent to any
        of the expected translations.
    *** Your eighteenth answer, "", isn't well formed.
    *** Your nineteenth sentence, "Large(a)->Small(d)", is not equivalent to
        any of the expected translations.
    *** Your twentieth sentence, "Large(a)->Large(e)", is not equivalent to
        any of the expected translations.
```

Fig. 4. Example feedback from the Grade Grinder: A translation exercise with errors

Domain Language. While the majority of the exercises in LPL use the language of the blocks world used in Figure 2, eight translation exercises use one of two other languages. In particular we have a language involving the care and feeding of various pets by their associated people. In this language, it is possible to give a translation for sentences like *Max fed Pris at 2:00*. This language is used in six of the translation exercises. The third language is used in only two exercises and is used to make claims about numbers, such as *There is a number which is both even and prime*.

Supporting and Additional Tasks. Each of the exercises in the pet and number languages require only the translation of sentences from NL into FOL. However, the use of the Tarski's World application provides scope for variety in the blocks language tasks. For example, some exercises call for students to complete their translations while looking at a world in which the English sentences are true; some call for them to verify the plausibility of their answers by examining a range of worlds in which the sentences have different truth values; and yet others call for the students to build a world making all of the English sentences true *de novo*. These alternatives represent a range of exercises in which the *agency* of the student varies. The act of constructing, from scratch, a blocks world that is consistent with a list of sentences (such as Example 7.15) requires more engagement and 'deeper' processing than one in which the student checks the truth of a sentence against a pre-fabricated diagram (such as Example 7.1). The effect of this variety in agency is one of many possible analyses that could be carried out using this corpus.

Figure 5 lists the different translation exercises and their characteristics. The 'Language' column indicates the target language, which is full FOL unless otherwise noted. In the exercises involving the blocks world language, the different kinds of agency that the students have are indicated. **Looking at world** indicates that students are instructed to look at a world in which the sentences are true as they translate the sentences, while **with world check** means that students are instructed to check their translations in specific worlds after the exercise is completed. **With world construction** indicates that students are required to construct

Exercise	Sentences	Language	Supporting Tasks
1.4	12	blocks (atoms)	
3.20	10	blocks (Boolean)	indirect + looking at world
3.21	12	blocks (Boolean)	with world check in next exercise
7.11	10	blocks (Propositional)	indirect + looking at world
7.12	20	blocks (Propositional)	with world check in next exercise
7.15	12	blocks (Propositional)	with world construction
9.12, 11.4, 14.4	10, 8, 7	blocks	indirect + looking at world
9.16	16	blocks	one existential + with world check
9.17	15	blocks	one universal + with world check
9.18, 11.14, 11.40, 14.28	5, 2, 11, 5	blocks	looking at world, with world check
11.16	10	blocks	skeleton translation given + with world check
11.17, 11.18, 11.19, 14.3	10, 5, 5, 5	blocks	with world check
11.20, 11.39	12, 6	blocks	looking at world
14.6	11	blocks	incomplete information
14.8	2	blocks	
14.27	2	blocks	with world construction
1.9	6	pet (atoms)	
3.23	6	pet (Boolean)	
7.18	5	pet (Propositional)	
9.19, 11.21, 11.41	10, 10, 5	pet	
9.13, 9.25	5, 5	number	

Fig. 5. Exercises involving English sentences (N=33)

(and submit) a world in which their sentences are true. **Incomplete information** means that not all relevant aspects of the world that they are looking at can be seen (e.g., a block may be obscured by a larger one).

The remaining annotations reflect other information given to the student. **Indirect** indicates that translations are given in the form 'Notice that all the cubes are universal. Translate this'. In the exercises marked **with one existential/universal** students are told that their translations have the specified form, while **skeleton translation given** indicates that students are given a partial translation that they must complete.

4. THE DATA IN THE TRANSLATIONS SUBCORPUS

The Translations Subcorpus represents all of the solutions to translation exercises submitted in the period 2001–2010. Translation exercises have in common that some number of sentences must be translated from NL into FOL. As noted above, we refer to the submission of a single answer to the translation of a sentence as a **translation act**; the corpus records a row of data for each translation act consisting of:

Unique ID. The unique identifier of this translation act (an integer).

Submission ID. The unique identifier of the submission in which this act occurs (an integer).

Subject ID. The unique identifier of the subject performing this act (an integer).

Instructor ID. The unique identifier of the instructor to whom this submission was copied (an integer). This field can be empty if the submission was not copied to an instructor.

Task. An indication of the task to which this is a response (for example, 'Exercise 1.4, Sentence 7').

Status. One of the values **correct**, **incorrect**, **ill-formed**, **not-a-sentence**, **undetermined**, **missing** (explained further below).

Answer. The text of the subject's answer (a string).

Canonical. The canonicalized text of the subject's answer (a string), where canonicalization simply involves removing whitespace from the answer, so that we can recognize answers which differ only in the use of whitespace.

Field	A Correct Act	An Incorrect Act
ID	7982509	7982763
Submission ID	3808583	4172630
Subject ID	68114	68114
Instructor ID	NULL	NULL
Task	7.12.1	7.12.15
Status	correct	incorrect
answer	$\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,d)}$	$\mathsf{Cube(a)} \rightarrow \mathsf{Large(a)}$
canonical	$\mathsf{Tet(a)} \rightarrow \mathsf{FrontOf(a,d)}$	$\mathsf{Cube(a)} \rightarrow \mathsf{Large(a)}$
Timestamp	2009-05-02 14:01:24	2009-05-02 14:49:32
File Timestamps	C1241297735665D1241298049184	*suppressed—see text*

Fig. 6. Example data for two translation acts from the corpus

	Correct	Incorrect	Missing	Ill-formed	Non-sentence	Undetermined	Total
First Submission	3,260,979	604,965	481,851	233,605	19,378	45,085	4,645,863
	70%	13%	10%	5%	0.4%	0.9%	
All Submissions	17,254,818	1,805,268	481,851	843,183	58,532	245,055	20,688,707
	83.40%	8.73%	2.33%	4.08%	0.28%	1.18%	

Fig. 7. Total submitted translation acts, classified by status

Timestamp. The time at which the submission was made.

File Timestamps. An indication of timing data concerning the file in which this act appears (explained further below).

Corpus data for two translation acts are shown in Figure 6. Each is an answer to one task within Exercise 7.12 (see Figure 2); the first data column shows a correct answer for Sentence 7.12.1, and the second represents an incorrect answer for Sentence 7.12.15.

The different **Status** values indicate different conditions that can occur when the student's submitted sentence is judged against the gold-standard answer. In addition to **correct** and **incorrect**, a solution may be **ill-formed**, indicating that the solution is not syntactically correct; **not-a-sentence**, indicating a well-formed FOL expression which does not express a claim (the closest analog in NL is a sentence with an unresolved anaphor); or **undetermined**, indicating that the Grade Grinder could not determine whether the submitted answer was correct. Finally, a solution can be **missing**. Because translations are packaged together into submissions of solutions for an exercise which contains multiple translation tasks, we code a solution as **missing** if the subject submitted translations for some, but not all of, the sentences in the exercise. A status of **missing** therefore represents a missed opportunity to submit a solution to accompany others that were submitted.

File Timestamps are an integral part of the Grade Grinder system, and record the times of save and read operations on the submissions file being constructed on the user's desktop. Each time a student opens or saves a file, a timestamp for this operation is added to a collection which is stored in the file. The collection of timestamps serves as a 'fingerprint' for the file, which allows the Grade Grinder to detect the sharing of files between students. Since these timestamps are accurate to the millisecond, it is extremely unlikely that files constructed independently will share any timestamps, and so two students submitting files whose timestamps are the same have likely shared the file. This fingerprinting mechanism is similar to the more familiar checksum algorithms which are often used to fingerprint files; the difference here is that the timestamp fingerprints are not dependent on the content of the file. This is important since some LPL exercise have a unique solution: consequently, arrival at the same content should not be considered evidence of sharing of a file.

Note that this timestamp data can be used to measure the amount of time that subjects spent considering their answers at a more fine-grained level than is indicated by the time between submissions. In the case of the first answer in Figure 6, the timestamp indicates that this file was opened (the segment beginning with C) and then saved (the segment beginning with D) about five minutes later (313,519ms being the precise difference between the two numbers). The timestamp data for the second answer contains fifteen segments, and so has been suppressed here because it is too large to display.

5. SOME SUMMARY DATA

The corpus contains a total of 4,645,563 initial submissions of translation acts by students, with 604,965 (13%) considered to be in error by the Grade Grinder. The breakdown of these initial submissions as provided by the Grade Grinder is shown in the upper half of Figure 7.

In fact, however, these numbers form a lower bound on the number of translation acts in the corpus. As noted earlier, a typical interaction with the Grade Grinder consists in a sequence of submissions, each of which may contain many translation acts. Initially, some of the translations in the submission will be correct and others incorrect. In each subsequent submission, some of the incorrect sentences will be corrected, while the correct sentences will be resubmitted; finally, the student may verify that all sentences are correct, and the student will likely then resubmit the complete set copied to their instructor. We therefore store multiple instances of the same translation acts.

The same phenomenon impacts on incorrect translation acts. If a student has made a mistake in both Sentence n and Sentence $n+1$, a common behavior is to repeat the submission first with a correction for Sentence n, but leaving the incorrect translation of Sentence $n+1$ unmodified from the previous submission, only returning to this once a correct answer for Sentence n has been achieved. This results in multiple instances of the same incorrect translation act. However, it is important to observe that in some cases these resubmitted incorrect answers may reflect deliberate acts, and so the real number of intended translation acts in the corpus may in fact be larger than our initial counts suggest. We provide all translation acts in the distributed corpus, with the corresponding counts shown in the lower half of Figure 7. The distributed corpus thus contains a total of 20,688,707 translation acts; this opens the door to additional analyses that would not be possible if only first submissions were available.

Note that we count as errors only those translations that are assessed by the Grade Grinder as definitely incorrect. Expressions which are offered as translations but which are not well-formed expressions of FOL, and those which are well-formed but not sentences, are counted separately. Of course, these expressions are really different kind of errors, and may serve to shed light on student behavior in other ways.

Among the translation exercises, the sentences most commonly mistranslated on the student's first attempt are shown in Figure 8. In this figure, the column headed **N** represents the total number of translation acts concerning this sentence, while the column headed **error/N** is the proportion of these acts that are marked as incorrect. The column headed **Count** applies to the distinct incorrect sentences, and indicates the number of translation acts that result in this answer.

6. POTENTIAL ANALYSES OF THE CORPUS

We conclude by outlining a number of ways in which the Translations Subcorpus can be analysed.

Sentence Features. What features of sentences are particularly difficult for all students (in the aggregate) to translate? We report on work of this type in [Barker-Plummer et al. 2011]. We categorized the sentences according to whether they contained shape, size and spatial predicates, and then examined the error rates for for eight resulting types of sentences. Sentences that mix shape and spatial predicates, and size and spatial predicates are each harder to translate than sentences that contain all three kinds of predicates.

Task		Answer	N	Error/N	Count
11.39.4		Every small cube is in back of a particular large cube	3520	69.0%	
11.39.4	**Correct**	$\exists x\,(\mathsf{Large}(x) \wedge \mathsf{Cube}(x) \wedge \forall y\,((\mathsf{Small}(y) \wedge \mathsf{Cube}(y)) \rightarrow \mathsf{BackOf}(y,x)))$			
	Incorrect	$\forall x\,((\mathsf{Cube}(x) \wedge \mathsf{Small}(x)) \rightarrow \exists y\,(\mathsf{Cube}(y) \wedge \mathsf{Large}(y) \wedge \mathsf{BackOf}(x,y)))$			818
	Incorrect	$\forall x\,((\mathsf{Small}(x) \wedge \mathsf{Cube}(x)) \rightarrow \exists y\,(\mathsf{Large}(y) \wedge \mathsf{Cube}(y) \wedge \mathsf{BackOf}(x,y)))$			420
	Incorrect	$\forall x\,((\mathsf{Cube}(x) \wedge \mathsf{Small}(x)) \rightarrow \exists y\,(\mathsf{Large}(y) \wedge \mathsf{Cube}(y) \wedge \mathsf{BackOf}(x,y)))$			281
	Incorrect	$\forall x\,\exists y\,((\mathsf{Cube}(x) \wedge \mathsf{Small}(x)) \rightarrow (\mathsf{Cube}(y) \wedge \mathsf{Large}(y) \wedge \mathsf{BackOf}(x,y)))$			207
	Incorrect	$\forall x\,((\mathsf{Small}(x) \wedge \mathsf{Cube}(x)) \rightarrow \exists y\,(\mathsf{Cube}(y) \wedge \mathsf{Large}(y) \wedge \mathsf{BackOf}(x,y)))$			164
11.20.1		Nothing to the left of **a** is larger than anything to the left of **b**	9101	54.9%	
11.20.1	**Correct**	$\neg\exists x\,(\mathsf{LeftOf}(x,a) \wedge \forall y\,(\mathsf{LeftOf}(y,b) \rightarrow \mathsf{Larger}(x,y)))$			
	Incorrect	$\forall x\,\forall y\,((\mathsf{LeftOf}(x,a) \wedge \mathsf{LeftOf}(y,b)) \rightarrow \neg\mathsf{Larger}(x,y))$			941
	Incorrect	$\forall x\,(\mathsf{LeftOf}(x,a) \rightarrow \forall y\,(\mathsf{LeftOf}(y,b) \rightarrow \neg\mathsf{Larger}(x,y)))$			913
	Incorrect	$\neg\exists x\,(\mathsf{LeftOf}(x,a) \wedge \forall y\,(\mathsf{LeftOf}(y,b) \wedge \mathsf{Larger}(x,y)))$			582
	Incorrect	$\forall x\,\forall y\,((\mathsf{LeftOf}(x,a) \wedge \mathsf{LeftOf}(y,b)) \rightarrow \neg\mathsf{Larger}(x,y))$			406
	Incorrect	$\forall x\,(\mathsf{LeftOf}(x,b) \rightarrow \neg\exists y\,(\mathsf{LeftOf}(y,a) \wedge \mathsf{Larger}(y,x)))$			307
3.21.5		Neither **e** nor **a** is to the right of **c** and to the left of **b**	34608	54.4%	
3.21.5	**Correct**	$\neg(\mathsf{RightOf}(e,c) \wedge \mathsf{LeftOf}(e,b)) \wedge \neg(\mathsf{RightOf}(a,c) \wedge \mathsf{LeftOf}(a,b))$			
	Incorrect	$\neg(\mathsf{RightOf}(e,c) \wedge \mathsf{RightOf}(a,c)) \wedge \neg(\mathsf{LeftOf}(e,b) \wedge \mathsf{LeftOf}(a,b))$			4681
	Incorrect	$\neg\mathsf{RightOf}(e,c) \wedge \neg\mathsf{RightOf}(a,c) \wedge \neg\mathsf{LeftOf}(e,b) \wedge \neg\mathsf{LeftOf}(a,b)$			1777
	Incorrect	$\neg(\mathsf{RightOf}(e,c) \wedge \mathsf{LeftOf}(e,b)) \vee \neg(\mathsf{RightOf}(a,c) \wedge \mathsf{LeftOf}(a,b))$			1678
	Incorrect	$\neg(\mathsf{RightOf}(e,c) \vee \mathsf{RightOf}(a,c)) \wedge \neg(\mathsf{LeftOf}(e,b) \vee \mathsf{LeftOf}(a,b))$			1569
	Incorrect	$\neg(\mathsf{RightOf}(e,c) \wedge \mathsf{RightOf}(a,c) \wedge \mathsf{LeftOf}(e,b) \wedge \mathsf{LeftOf}(a,b))$			1345
3.23.5		2:00pm is between 1:55pm and 2:05pm	14747	50.4%	
3.23.5	**Correct**	$1:55 < 2:00 \wedge 2:00 < 2:05$			
	Incorrect	$\mathsf{Between}(2:00, 1:55, 2:05)$			14546
	Incorrect	$\mathsf{Between}(1:55, 2:00, 2:05)$			319
	Incorrect	$\mathsf{Between}(2:00, 2:05, 1:55)$			178
	Incorrect	$\mathsf{Between}(2, 1:55, 2:05)$			133
	Incorrect	$2:00 < 2:05$			91
11.40.3		There is a dodecahedron unless there are at least two large objects	3887	48.7%	
11.40.3	**Correct**	$\neg\exists x\,\exists y\,(x \neq y \wedge \mathsf{Large}(x) \wedge \mathsf{Large}(y)) \rightarrow \exists z\,\mathsf{Dodec}(z)$			
	Incorrect	$\exists x\,\exists y\,(\mathsf{Large}(x) \wedge \mathsf{Large}(y) \wedge x \neq y) \rightarrow \neg\exists z\,\mathsf{Dodec}(z)$			84
	Incorrect	$\exists x\,\exists y\,((\mathsf{Large}(x) \wedge \mathsf{Large}(y) \wedge x \neq y) \rightarrow \neg\exists\,z\mathsf{Dodec}(z))$			67
	Incorrect	$\exists\,x\exists\,y\exists z\,(\mathsf{Dodec}(x) \rightarrow \neg(\mathsf{Large}(y) \wedge \mathsf{Large}(z) \wedge y \neq z))$			54
	Incorrect	$\exists x\,\exists y\,((\mathsf{Large}(x) \wedge \mathsf{Large}(y) \wedge x \neq y) \rightarrow \exists z\,(\mathsf{Dodec}(z) \wedge z \neq x \wedge z \neq y))$			48
	Incorrect	$\forall x\,\forall y\,((\mathsf{Large}(x) \wedge \mathsf{Large}(y) \wedge x \neq y) \rightarrow \neg\exists z\,\mathsf{Dodec}(z))$			46

Fig. 8. The top five erroneous answers to the each of the five most error-prone tasks

Error Typology. Can the errors that students make in their translations be categorized according to type? In [Barker-Plummer et al. 2008] we examined the most frequent errors in the solution of Exercise 7.12, and discovered that the failure to distinguish between the conditional and biconditional was a significant source of error. Another significant source of error appears to be an expectation that names will appear in contiguous alphabetical order in a sentence (we call these 'gappy' sentences); so, a sentence like '**a** is between **b** and **d**' is frequently mistranslated with **c** in place of **d**.

Response to Errors. How do subjects go about finding solutions when their initial attempt is incorrect? We can ask whether the difficulty of repair correlates with the subject, the sentence or with the particular error that was initially made. We have carried out preliminary work [Barker-Plummer et al. 2009] investigating the differences between, on the one hand, translation tasks which are difficult to get right initially but which

are easy to recover from, and on the other hand, those which are perhaps less error-prone, but hard to repair. We think both aspects of the task contribute to the 'difficulty' of a task.

Exercise-Level Strategies. There is potential in the corpus for examining strategies that the students adopt when they make multiple errors. Some students appear to attempt to fix all of their incorrect sentences, and others proceed one at a time. These strategies might correlate with success. We can detect differences between these strategies by looking at the sequence of submissions that occurs after the initial submission. In some cases only one sentence will be modified in each subsequent submission; in others many may be altered.

Modality Heterogeneity of Task. Exercises differ in the extent to which they are linguistically and graphically heterogeneous. Some require translation from NL sentences to FOL, whereas others require translation followed by blocks world diagram building. In [Cox et al. 2008], we compared students' constructed diagrammatic representations of information expressed in NL sentences to their FOL translations, and determined that the error patterns differed in their graphical versus their FOL translations.

Agency in the Task. As discussed in Section 3, translation tasks vary in the degree of agency they require on the part of the student. Using the corpus it would be possible to analyze variability in student performance with agency, to see if these adjunct tasks have an effect on translation accuracy.

Time Course. The timestamp information in the corpus makes it possible to ask how much time students spend (re)considering their answers: does the bulk of time go to particular tasks, or is it evenly distributed?

7. CONCLUSION

With the first release of this corpus, we invite colleagues to exploit its potential for educational data mining. Our hope is that further analyses will provide additional insights into student cognition in the difficult domain of logic, and that findings will inform improved educational practice in logic teaching. In our own work, we aim to (1) enrich the feedback that Grade Grinder provides to students, (2) investigate task agency effects upon learning outcomes, and (3) identify evidence-based improvements to the logic curriculum.

REFERENCES

Barker-Plummer, D., Cox, R., and Dale, R. 2009. Dimensions of difficulty in translating natural language into first order logic. In *Second International Conference on Educational Data Mining*. Cordoba Spain.

Barker-Plummer, D., Cox, R., and Dale, R. 2011. Student translations of natural language into logic: The Grade Grinder corpus release 1.0. Technical Report OP-TR-01. Available from `openproof.stanford.edu`.

Barker-Plummer, D., Cox, R., Dale, R., and Etchemendy, J. 2008. An empirical study of errors in translating natural language into logic. In *Proceedings of the 30th Annual Cognitive Science Society Conference*, V. Sloutsky, B. Love, and K. McRae, Eds. Lawrence Erlbaum Associates.

Barker-Plummer, D., Dale, R., and Cox, R. 2011. Impedance effects of visual and spatial content upon language-to-logic translation accuracy. In *Proceedings of the 32nd Annual Cognitive Science Society Conference*, C. Hoelscher, T. F. Shipley, and L. Carlson, Eds. Lawrence Erlbaum Associates.

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

Corbett, A. T. and Anderson, J. R. 1989. Feedback timing and student control in the lisp intelligent tutoring system. In *Proceedings of the Fourth International Conference on Artificial Intelligence and Education*, D. Bierman, J. Brueker, and J. Sandberg, Eds. IOS Press Amsterdam Netherlands.

Cox, R., Dale, R., Etchemendy, J., and Barker-Plummer, D. 2008. Graphical revelations: Comparing students' translation errors in graphics and logic. In *Proceedings of the Fifth International Conference on the Theory and Application of Diagrams*, G. Stapleton, J. Howse, and J. Lee, Eds. Lecture Notes in Computer Science LNAI 5223, Berlin: Springer Verlag.

Koedinger, K., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., and Stamper, J. 2010. A data repository for the EDM community: The PSLC datashop. In *Handbook of Educational Data Mining*, C. Romero, S. Ventura, M. Pechenizky, and R. Baker, Eds. CRC Press Boca Raton Florida.