The pragmatics of margin comments: An empirical study

Debora Field, Stephen Pulman
Dept Computer Science
University of Oxford
Oxford OX1 3QD, UK
firstname.lastname@cs.ox.ac.uk

Denise Whitelock
Institute of Educational Technology
The Open University
Milton Keynes MK7 6AA, UK
denise.whitelock@open.ac.uk

Abstract

This paper describes the design and rationale behind a classification scheme for English margin comments. The scheme's design was informed by pragmatics and pedagogy theory, and by observations made from a corpus of 24,387 margin comments from assessed university assignments. The purpose of the scheme is to computationally explore content and form relationships between margin comments and the passages to which they point. The process of designing the scheme resulted in the conclusion that margin comments require more work to understand than utterances do, and that they are more prone to being misunderstood.

1 Introduction

We have a collection of 24,387 real margin comments, expressed in English, which we want to exploit through machine learning in order to inform the design of an automatic margin comments generator. The corpus margin comments were added by humans to a corpus of real assessed university assignments. The assignments were argumentative essays submitted towards a Master's degree in Education.

We have designed a margin comment classification scheme which classifies natural language (NL) margin comments without reference to the essay parts to which they point. High inter-annotator agreement scores have been achieved for the scheme. We plan to use the scheme to look for relationships between the corpus comments and the essay parts to which they point.

This paper is about the classification scheme's design, including what led to the design decisions, which were informed by examination of the margin comments, the assignments corpus, and consideration of key ideas in pragmatics and pedagogy. A feature of margin comments that became clear during the design process, and that influenced the design, is that margin comments are harder to understand and are more prone to being misunderstood than conversational utterances.

2 What are the corpus comments like?

The design of the classification scheme is based on answers we sought to three core questions:

- What are the margin comments like?
- What are they 'doing'?
- How do they get their messages across?

A margin comment is a message written or typed by an assessor and positioned in the 'margin' of a piece of text produced by a learner. Most margin comments graphically point to a part of the learner text, and the message content of a margin comment typically concerns the text part to which the comment points. The margin comments in our corpus had been added to word-processed assignments using a digital commenting tool.

To gain a first impression of what the corpus margin comments were like, we carried out some frequency counts and from these derived a set of simple pattern-matching rules for clustering similar comments—143 complex regular expressions to match the start of a comment. Most of the rules invoked one or more of 13 regex groups. Each group was a disjunction of strings (*e.g.,* 29 'negative' verb disjuncts). Each comment was typed on the basis of its first sentence only, on the grounds that any subsequent sentences were most likely elaborations on the first (based on manual scrutiny of hundreds of comments.) Probable comment-initial filler words were skipped. The clustering rules assigned a type to 90.9% of the comments. The following subsections describe some of the results.

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications , pages 43–53,
Baltimore, Maryland USA, June 26, 2014.

2.1 Positive-sounding

Expressions that are positive-sounding in general (*e.g.,* 'good', freq. 5,177) and positive with respect to essay writing (*e.g.,* 'interesting', freq. 954) were very common.[1] There were 9,272 occurrences of a positive-sounding adjective. In contrast, there were 551 occurrences of a negative-sounding adjective, the top 3 being 'difficult' (freq. 133), 'missing' (123), 'informal' (90). A large proportion of positive-sounding comments were descriptions. For example, 3,151 comments (12.9%) began with 'good'.

2.2 Missing, unnecessary, or inappropriate

3,351 comments expressed the idea that something was missing from the essay that marker M thought should have been present (1a). 574 comments expressed the idea that something was present in the essay that M thought should not have been (1b). 2,069 comments expressed the idea that something that was present in the essay that M thought should have been different in some way (1c).[2]

(1) a. Could you have developed this?
    b. I would not leave a space.
    c. Another long quote

2.3 Confusion and apparent uncertainty

1,119 comments expressed confusion or apparent uncertainty. Many confusion expressions concerned M's understanding. There were 1,232 expressions concerned with comprehensibility. Many uncertainty expressions concerned M's agreement or understanding. There were 1,193 expressions concerned with agreement.

2.4 Questions

4,307 comments (17.6%) ended in a question mark and 1,109 comments began with a WH question word. 1,119 comments were polar questions.

2.5 Parts of instructions

6,169 expressions looked like parts of instructions or polite suggestions, the top 3 being 'you might' (freq. 882), 'you need' (693) and 'explain' (332).

2.6 Adversative conjunctions

There were 2,237 occurrences of 'but', 283 of 'although', 127 of 'however', typically used in the corpus to present contrasting or opposing opinion.

2.7 Non-sentential

The distribution of comment lengths is heavily skewed towards short comments (Figure 1).[3] Just under 9.5 % of comments have 11 characters or fewer. The top 3 most frequent comment lengths were 10 characters (freq. 430), 4 characters (freq. 358) and 1 character (freq. 316).

Scrutiny of many short comments revealed that non-sentential comments are the main reason for the brevity. These include elliptical comments (2a), fragments (2b), and other non-sentential expressions such as exclamations (2c) and short directives (Klein, 1985; Merchant, 2004) (2d).

(2) a. Why not?
    b. Good point
    c. What a good idea.
    d. Reference

Very short corpus comments that are complete sentences are rare (set 3).

(3) a. Avoid jargon
    b. This is unclear.

2.8 Politeness

There are 3,996 occurrences of terms typically used to soften the impact of a criticism or make an instruction sound like a suggestion (hereon 'softeners'), including 'perhaps' (freq. 863), 'rather' (422), and 'a little' (381). There are also 7,287 occurrences of conditional auxiliary verbs (including many non-modal uses of 'would'), which are typically used to make polite suggestions.

2.9 Informality

There are 3,818 contractions, including "don't" (freq. 568), "I'm" (370), "you're" (138). Filler words were also common. 444 comments began with 'ok' (a range of spellings), and 1109 comments began with 'yes' (some of these express agreement, but most are fillers).

2.10 Skills

We noticed 4 large groups of terms relating to particular skills. Table 1 shows each group, the number of occurrences of terms from that group, an example term from that group, and the number of occurrences of the example. Category 'presentation' includes matters relating to the presentation of English, such as spelling, grammar, formatting, and style.

[1] All quoted example terms are case-insensitive.

[2] All examples in the paper are real, whole comments from the corpus, apart from examples that are prefixed with a ' ^ ', which are interpretations. Punctuation, spelling, capitalisation, *etc.*in the examples are faithfully reproduced.

[3] The inset in Figure 1 is the main figure presented on log-log scale axes.

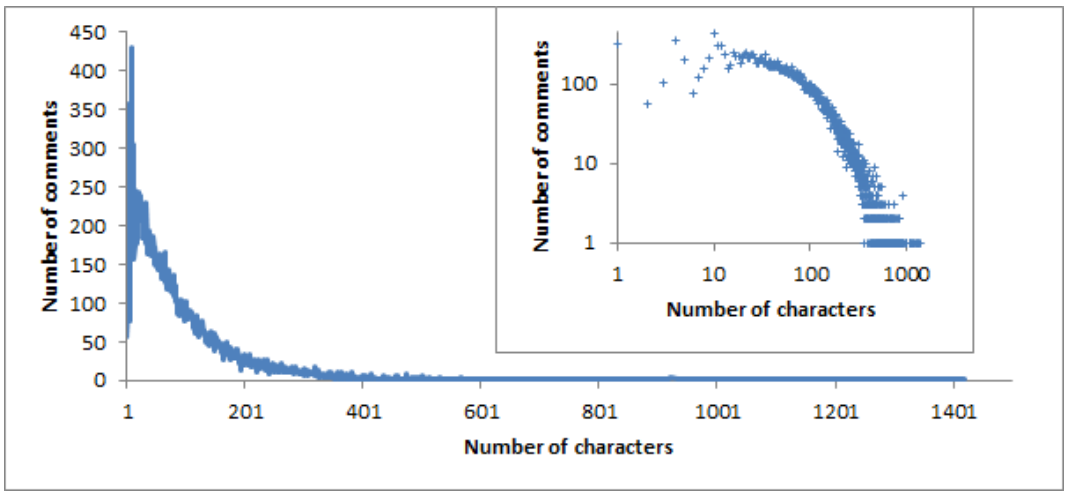

Figure 1: Distribution of comment lengths

Grouping	Freq.	Example	Freq.
Argument	14705	'argument'	817
Referencing	6657	'reference'	1322
Essay structure	5243	'section'	614
Presentation	2613	'sentence'	428

Table 1: Skills-related terms in comments corpus

2.11 Are margin comments conversation?

The corpus investigations revealed frequent use of phenomena common in speech: non-sentential expressions, contractions, politeness devices, softeners, and fillers. This led us to consider whether a dialogue act taxonomy such as DIT (Bunt, 1990) or DAMSL (Core and Allen, 1997) might be suitable for typing margin comments.

Many pedagogy papers have argued or assumed that margin comments are or are like a conversation. Straub (1996) reviewed a number of contemporary papers, including (Ziv, 1984; Danis, 1987; Lindemann, 1987; Anson, 1989) to explore the question: "what does it mean to treat teacher commentary as a dialogue?" (Straub, 1996, p. 375). Straub concluded that margin comments are not conversational utterances, either real or imaginary, and that what pedagogy scholars were referring to was the informal style of comments. Informal language was becoming popular as a result of the movement away from 'teaching product' to 'teaching process', which encouraged the expression of empathy with the learner, because it was thought this would make teacher comments more likely to be read and acted upon (Hairston, 1982).

From linguistics, Schegloff (1999), tackles the problem of whether there is such a thing as 'ordinary conversation'. He first defines 'talk-in-interaction' (p. 406), which includes speech spoken with the intention of communicating messages to some audience. Next Schegloff talks about 'speech exchange systems' (citing (Sacks et al., 1974)), which are "organizational formats for talk-in-interaction" (Schegloff, 1999, p. 407), including the lecture format, classroom discourse, courts-in-session, meetings, debates, *etc.* Margin comments arguably qualify as a speech exchange system, even though they are written, not spoken. But Schegloff's definition of 'ordinary conversation' arguably excludes margin comments on the grounds that they *don't* involve "generic aspects of talking-in-interaction such as turn-taking, sequence organization, repair organization, overall structural organization" (p. 413), and on the grounds that they *are* "subject to functionally specific or context-specific restrictions" (p. 407).

Having considered relevant literature, we concluded that margin comments are not conversation, principally on the grounds that there is no turn taking—only the marker M gets the opportunity to 'speak' and only the comment's addressee A gets the opportunity to 'hear'. Whilst there is common ground (Stalnaker, 1972; Thomason, 1990) and accommodation (Clark and Haviland, 1974; Lewis, 1979; Kamp, 1981) there is no turn taking and therefore no grounding (Clark and Schaefer, 1989). M presents utterances, and the constraints of the context demand that A must accept the evidence. Consequently, if A misunderstands M's intended message, there is no mechanism to enable M or A to discover that A has mis-

understood the message; and if A is confused by M's comment, there is no opportunity for A to ask M for clarification.

We concluded that dialogue acts were an inappropriate classification scheme for margin comments, because the conditions for human-to-human dialogue do not apply.

3 What are margin comments 'doing'?

If dialogue acts are inappropriate, what kinds of things *are* NL margin comments 'doing'? Consider WH questions (4).

(4) Why bold?

When M asks a WH question in a margin comment, M is not desiring or expecting A to supply the requested information to M. The Addressee A of a NL margin comment will never take a turn in response to that comment. This is something of which A and marker M are both mutually aware before the comments are written by M, and it has important repercussions with respect to M's intentions. Consider also imperatives (5).

(5) Explain what they do.

5 looks like an instruction, but cannot be. The corpus comments were added to the final, submitted versions of assessed assignments. There was no desire or expectation on M's part that A would revise the essay in response to M's comments.

M must have been desiring *something* by these comments (otherwise there would be no comments), but that something is not what one might expect given their linguistic surface forms. This suggests that margin comments are like indirect speech acts (Searle, 1969; Searle and Vanderveken, 1985)—acts which have an *apparent* function that is distinct from what the comment is *really* 'doing' (Austin, 1962). We would argue that, for the evaluative comments in the corpus (which are the vast majority), the thing the comments are doing is this: **to communicate M's opinion to A about the essay part to which the comment pointed**.

This conclusion is not surprising. NL margin comments are doing what all margin comments are doing, it seems, including non-NL coded comment schemes. Why this conclusion *seems* surprising is that margin comments do not look like expressions of opinion about weaknesses and strengths. Instead they look like excerpts from friendly, informal conversations. The informality is, however, masking the principal messages of the comments, which are evaluative ones.

4 How do NL margin comments express whether the essay met the standard?

Having decided what NL margin comments are doing, we reasoned that M's opinion expressed by a comment must have two aspects, on the grounds that they do not just point to essay parts, they contain messages. The two aspects are: (1) ***Whether* or not essay part P to which a comment points attained the required standard**; (2) ***How* P attained (or did not attain) the required standard**. The required standard is a standard defined by some set of principles or instructions of which M and A are typically mutually aware.

We observed that the semantics of very few corpus comments communicated a message approaching 'This essay part has failed to achieved the agreed standard'. Set 6 shows two of them.

(6) a. Something's wrong or missing here...
b. Two line sentences is not enough to get the maximum 30% marks for this section

For the vast majority of comments, ***whether* essay part P attained the required standard was communicated implicitly by the use of certain types of words and syntactic structures**. To convey attainment or surpassing of the standard, positive-sounding adjectives were used extensively (section 2), also positive-sounding adverbs, and terms of liking, agreement, and understanding. A much wider variety of techniques was used to convey *failure* to attain the standard, including negative-sounding verbs (*e.g.*, 'contradict'), negative-sounding adjectives (*e.g.,* 'inappropriate'), lone noun phrases (*e.g.,* 'brackets'), questions, instructions, polite suggestions, notifications of marker edits, referrals to authoritative sources, and assertions of uncertainty, confusion, doubt, disagreement, and non-understanding.

Addressee A's understanding of whether essay part P had attained the required standard would therefore have depended on A's being able to correctly interpret the semantics of the comment. For non-native speakers of English, this may have presented a problem.[4] Since many corpus comments

[4]The corpus assignments were towards a distance-learning degree course, and many of the students are likely to have been non-native speakers of English.

constitute a lone modified noun phrase, and since the meanings of everyday adjectives change depending on what they are modifying, it may have been difficult for A to tell whether a comment was a criticism or a commendation (set 7).

(7) a. A very long sentence.
b. Very strong supporting quote.
c. A strong argument
d. A big assumption

Note that the way we decide whether these are criticisms or commendations is by considering the type of entity the adjective is modifying. We know quotes should be strong, so 7b must be a commendation. We know assumptions should not be big, so 7d must be a criticism. This means that, in addition to having a sensitivity to compositional semantics, the addressees of these comments would have needed to possess expert knowledge about what sentences, quotes, arguments, and assumptions should be like in order to be able to infer whether the essay part had met the standard.

Difficulties in understanding whether an essay part has met the standard are also caused by the use of non-sentential expressions (set 8).

(8) a. Reference
b. Colloquialism
c. No issues
d. No comma
e. No apostrophe

Which of the following interpretations (if any) applies to each of the set 8 comments?

(9) a. ˆ The named thing is missing
b. ˆ The spelling of the named thing is incorrect
c. ˆ The named thing is erroneously included
d. ˆ The named thing needs correcting
e. ˆ This part attains the required standard

In order to understand these comments, A has to inspect the passage to which the comment points to see whether it contains the object named by the comment. If it does, there may still be the possibility that it should be present, but that there is something wrong with it.

5 Scheme design: *Skill* targeted

We have considered what the corpus margin comments are doing, and the ways in which they express whether an essay part met the required standard. The way in which a comment conveys *h*ow the standard was or was not met is embodied by the comments classification scheme's design. The scheme has three layers, and here we consider the first. When M wrote a comment, M had in mind a good-essay-writing principle. Our classification scheme makes explicit the skill area of that essay-writing principle. Consider set 10.

(10) a. Why not?
b. Why bold?

To understand what these comments mean, we first need to know what M intended, which we have argued was to communicate to A whether and how the related essay part had reached an agreed standard. On that account, the comments (a) and (b) in 10 mean something like (a) and (b) in 11.

(11) a. ˆThe argument here would have been improved by including an explanation of why not.
b. ˆThe use of bold font here is questionable.

These are very different messages. One comment is alerting A to some missing argument, and the other is questioning A's use of different fonts. How do we know this, given that both comments have very similar syntactic structure?

Addressee A works out that these comments mean very different things by first identifying the skill area that the comment is targeting, and then considering what that skill area is *like*—in what ways it can be good or bad. To understand 10a, A needs to observe that essay part P contains a statement, and to infer that M is responding to the argument made by the statement. To understand 10b, A needs to observe that P contains some text in bold font, and to infer that M is questioning the use of the bold font. The difficulty here is that conversational-style comments do not make it explicit whether they are targeting content or form.

Concluding that the identification of a comment's target is often critical to understanding it, we defined 11categories for the scheme's 'targeted skill' layer. The corpus investigations (see 2.10) revealed four main skill areas targeted by comments:

- Referencing
  - Situating work in the relevant literature, referencing conventions
- Structuring Essays
  - Layout, scope, components
- Composing Argument
  - Content, quality, arguing techniques, comprehensibility
- Presenting English
  - Spelling, grammar, formatting, style

We made **Referencing** and **Structure** target categories in their own right. Owing to the high frequency of comments expressing confusion and comprehensibility (see 2.3) we made **Comprehensibility** a target category. Comments targeting the content of an argument, the quality of an argument (not including its comprehensibility), and arguing techniques are covered by target category **Argument**. We divide the skill area of presenting English into five subcategories: **Formatting**, **Grammar**, **Punctuation**, **Spelling**, **Style**.

An additional target category is **Context-Dependent**. This is assigned if an evaluative comment has very little information in it about what its targeted skill might be (set 12).

(12) a. Good [212 occurrances]
b. Avoid
c. Unfinished

The 11th target category, **Author**, is assigned to all comments which appear non-evaluative. These include, for example, casual observations, personal reminiscences, and expressions of gratitude.

6 Scheme design: marker's *Attitude*

Having concluded that each corpus comment was communicating M's opinion about an essay part, for the next layer in the scheme, we focused on opinion types. The investigation results revealed three common types (see section 2.2), which we named **Miss**, **Reject**, and **Condemn**. The attitudes do not involve the emotional connotations normally associated with these names in everyday communication. (Hereon we will refer to these as categories of attitude, rather than opinion.)

Having observed the large proportion of polar questions and expressions of uncertainty or doubt in the corpus (see 2.3 and 2.4), we decided to treat Miss, Reject, and Condemn as attitudes held by M with certainty, and to add another attitude **Doubt** to cover comments in which M called into question things that A had done, or in which M expressed some uncertainty or doubt.

- **Doubt**: "Why bold?"
  - M considers that something in the essay is of questionable value.

Since expressions of uncertainty are often used as softeners rather than to express actual uncertainty, it seemed inappropriate to treat apparent uncertainty as a qualifier (Bunt, 2011) of attitudes. If we treat it as a qualifier, it suggests that M was not sure about M's own opinion, rather than that the target of M's comment was questionable. Doubt is the attitude most applicable to the majority of polar questions in the corpus.

A further attitude, which is a sub-type of Condemn, is defined as **Dispute** (see section 2.6):

- **Dispute**: "Not necessarily."
  - M holds views that are in opposition to some proposition in the essay.

A further attitude **Commend** covers all comments that announce a 'strength' (see section 2.1):

- **Commend**: "Good"
  - M considers that something in the essay has attained or exceeded the required standard, or is pleasing or interesting to M.

Two further attitudes (**Refer** and **Exclaim**) are defined, which have a special characteristic.

- **Refer**: "Ditto."
  - M believes that A would benefit from reading a particular source.
- **Exclaim**: "Ah!"
  - M is surprised or shocked by something in the essay that M does not specify.

It is not possible to tell whether Refer comments are evaluative or not without reading the source to which M has referred the addressee. Similarly, it is impossible to tell whether Exclaim comments are evaluative or not, either from the comment or the essay part to which the comment points.

Two final attitudes—**Engage** and **Thank**—are reserved for non-evaluative comments, *i.e.,* comments whose target is Author.

- **Engage**: "I know how you feel."
  - M finds something about the essay or about A engaging. It appears that M has become engaged in a way that is more complex than liking or finding interesting.
- **Thank**: "Thanks"
  - M is grateful to A.

These attitudes are what we term 'solidarity' attitudes, in that we assume that they were made in order to engender positive feelings in A. Engage comments have a very wide variety of forms and topics, which we will not be attempting to analyse in the initial rounds of the machine learning trials. Thank comments are all expressions of gratitude.

7 Scheme design: *Linguistic Act*

The third layer of the categorisation scheme identifies what we are calling the 'linguistic act' of the comment. The acts are distinguished principally

by surface form and do not concern the evaluative (or non-evaluative) message that the comment is attempting to communicate.

We began with the three basic English sentence types: declarative, interrogative, imperative. We divided 'interrogative' into acts **WH Question** and **Polar Question**, as they have clearly distinguishable surface forms.

We also divided declarative comments into two acts: **Assertion** and **Description**. All margin comments, including interrogatives and imperatives, are by definition assertions of M's opinions, we have argued. The scheme's act Assertion is reserved for assertions of propositions in response to argument (13a, 13b) and explicit expressions concerning understanding (13c), agreement (13d), verification or certainty. Many assertions are subjective-sounding.

(13) a. That is impossible!
b. This is true of many other organisations
c. I don't understand
d. Not sure I agree!

Act Description is assigned to a comment which is a description of a (non-propositional) object in or quality of an essay part P or of an action that has been carried out by author A and that is evidenced by part P (set 14).

(14) a. Too many references.
b. Factors clearly articulated.
c. This is a very strong assertion

Splitting declaratives into acts Description and Assertion is a small step away from categorising linguistic acts according to syntax only. The move separates declarative comments which respond directly to propositional content from all other declarative comments.

We interpreted 'imperative' as linguistic act category **Instruction**. We treat the category loosely, allowing it to include comments that do not use the imperative form but that look like guidance on what should have been done (set 15).

(15) a. You should add a citation here.
b. I would not leave a space.
c. Ditto

All Instruction comments talk in a variety of ways about things that were not done but that should have been, whereas all Description comments (set 14) talk about what was actually done. This distinction is not too dissimilar to the distinction between imperatives and declaratives. That Instruction comments do not always have the imperative form is a repercussion of the informal conversational style of the comments.

A sixth 'dummy' linguistic act category is assigned to all comments with attitude Engage, because we will not be attempting to analyse those.

The linguistic act layer, then, categorises the comment's form, while the target and attitude layers categorise its meaning. The linguistic act accounts for what the comment is *apparently* doing (see section 3). The attitude and target account for what the comment is *really* doing. A stark difference between utterances and margin comments is that, to understand an utterance, hearer H does not have to work out what speaker S was really doing (Ramsay and Field, 2008); whereas to understand a margin comment, addressee A does have to work out what marker M was really doing.

8 Evaluation

We have demonstrated that the classification scheme can be deployed with high agreement levels between independent annotators. Agreement by two annotators was calculated for 313 sample comments that were annotated by each annotator independently. Annotator A designed the scheme over several months. Annotator B spent about 50 minutes learning the scheme (from no prior exposure to it). Annotator B took a mean average of 1.1 minutes to fully annotate each comment in the sample. Annotator A took a mean average of .49 minutes to fully annotate each comment.

The corpus comprised 1,408 essays submitted for 13 different assessed university Master's modules, the official word limits of which ranged from 500 to 4,000. The essays had been marked by 20 different markers. The number of essays marked by each marker varied. The mean average number of comments per essay per marker ranged from 4.83 to 47.00. To avoid potential bias towards the more prolific markers' styles, the same number of essays were randomly sampled for each marker (where possible), and approximately the same number of comments were randomly sampled from each of those essays.

Some tutors appear to prefer very short comments, some long. For some (but not all) of the tutors who marked essays of different lengths, there was a correlation between essay length and the number of margin comments. No analysis of linguistic style similarities across comments within

individual essays was carried out for this paper.

Inter-annotator agreement was calculated using Cohen's Kappa for each of the three layers of the scheme independently. 95% confidence intervals (CI) for test statistics were generated through 10,000 statistical bootstrappings of the annotated comments. The agreement coefficient for the attitude layer was 0.874 (95% CI, 0.831–0.914), for the target layer was 0.791 (0.734–0.844), and for the linguistic act layer was 0.822 (0.770–0.869). The percentage agreement across all three layers was 72.1% (67.0%–77.0%) (the percentage of comments for which both annotators were in agreement on all three layers). There were no occurrences of comments which both annotators deemed unclassifiable. One of the comments was deemed unclassifiable by one annotator.

The scheme has five attitude+target cross-layer dependencies (Engage+Author, Thank+Author, Refer+Context-Dependent, Exclaim+Context-Dependent, Dispute+Argument), and five target+act cross-layer dependencies (each of the same five pairs plus a linguistic act). We acknowledge that these might argue for a more complex agreement calculation. It is expected that some linguistic act categories are unlikely to combine with some attitude categories, though this requires empirical verification. A conservative estimate of the number of possible combinations of attitude, target, and act that we believe might be found in the corpus is 155 combinations. Additionally, some categories from a given layer appear to be more frequent than other categories from the same layer. We acknowledge, therefore, that a weighted coefficient method may be more suitable for calculating inter-annotator agreement.

9 Comparison with previous work

Now that the categorisation scheme has been described, we will discuss comparisons with previous work. Categorisation schemes have been devised or re-used in order to analyse written feedback, and discover where improvements might be made. The studies were principally interested in whether the marker was writing comments that would 'feed forward'. Measures for deciding whether a comment would feed forward tended to revolve around the power of a comment to motivate its addressee, or whether the comment contained explanatory text that would make it clear how to do things better in future. We have not found any feedback categorisation schemes primarily concerned with how opinion in comments is conveyed through the medium of NL.

Hyland (2001) designed a feedback classification scheme that was used to analyse the quality of feedback for a distance-learning language course. Hyland's scheme focused on targeted skills, affective aspects, and explicit pointers for future writing. Bales (1950) devised 12 categories for the purpose of analysing small group interactions, which were later applied to the analysis of margin comments by Whitelock *et al.* (2004). Bales' scheme focused on affective aspects (including solidarity, tension, antagonism), and pragmatics aspects (suggestions, opinions, disagreements, requests). Brown and Glover's (2006) scheme focused on skills, content, affective aspects, and feeding forward. They used their scheme to argue that the feedback in a particular corpus of comments was of limited value, because most of the comments did not aid learning or understanding (Brown et al., 2004). Nelson and Schunn (2009) wanted to identify conditions under which addressees of peer feedback might actually implement that feedback. Their categories focused on the linguistic features of comments (including summarisation, specificity, explanations) and affective issues. Perpignan (2003) viewed margin comments as part of dialogue and discussed the "intentions and interpretations of the exchange from both the teacher's and the learners' perspective" (p. 259). The work did not attempt to analyse the linguistic features of feedback.

The categorisation scheme with the strongest resemblance to ours was Ferris *et al.* (1997). The scheme viewed margin comments as having two 'phases': teacher's goal, and linguistic form (p. 163). The scheme has a very different interpretation of the intention of the marker from ours. It confuses marker intention with comment target. It implicitly recognises what we call marker attitude, but identifies only one (our Commend). It implicitly recognises the target of a comment but has only two target types ('form' and 'content').

10 Discussion

We have presented a classification scheme for margin comments which is based on observations of real data and on linguistics theory. The goal of the classification was to ultimately use machine learning to look for relationships between the mar-

gin comments and the essay parts to which they point so that we could design an automatic NL margin comments generator. The scheme therefore focuses on the linguistic aspects of margin comments: their form and meaning. It is designed to classify comments independently of the essay parts to which the comments point. This was to ensure that the comments in isolation could be classified to a useful level of agreement, and, in future work, to make it possible to investigate whether essay properties can be used to predict characteristics of margin comments. The 3-layered scheme enables the intended evaluative meanings of margin comments to be captured despite their conversational style, while also preserving linguistic information about that style (set 16).

(16) a. Could you have developed this?
   i. *Attitude:* Miss
   ii. *Target:* Argument
   iii. *Act:* Polar Question

   b. Why bold?
   i. *Attitude:* Doubt
   ii. *Target:* Formatting
   iii. *Act:* WH Question

   c. No issues
   i. *Attitude:* Commend
   ii. *Target:* Context-Dependent
   iii. *Act:* Assertion

The classification scheme is arguably a suitable scheme for all margin comments expressed in NL, with the proviso that the skills being targeted by the comments would need to be tailored to the document type if it were not an argumentative essay.

Details not discussed earlier include the following. (i) We use a skills precedence list to select a target for comments that are ambiguous over skill area. (ii) Prior to categorisation, each comment is segmented, and one 'principal segment' only is identified for categorisation. This is for 4 reasons. (1) Many comments begin with filler words (*e.g.,* 'yes', 'well', 'ok', 'hmm'); (2) Many begin with preambles (*e.g.,* 'minor point', 'Just one thought'); (3) Many use a commendation as a softener before delivering the main message; (4) Many contain more than one clause or sentence. We usually assume the first non-filler/non-preamble segment is the principal segment, and that any non-filler segments that follow are elaborations on the first segment. The exception to this is comments in which a commendation is used as a softener, in which case the segment that follows the softener becomes the principal segment.

High inter-annotator agreement scores have been achieved for the classification scheme. We have not yet annotated the whole corpus, but we intend to. We will also calculate inter-annotator agreement for a higher number of sampled comments, since the number of possible combinations of attitude, target, and act is so high (*circa* 155, see section 8). We may make small changes to the annotation scheme before doing any further annotation. In particular, we are considering dividing target Argument into two or three subcategories.

While designing the classification scheme, we have observed that, despite their conversational style—indeed because of it—understanding NL margin comments is harder than understanding conversational utterances. Although the essay part to which a comment points is a public object and therefore in M and A's views of the common ground, the aspect of the essay part that M is targeting usually remains unexpressed and private in M's mental state. A therefore has to do some inferencing to identify that aspect. In other words, A has to do some inferencing to fill in gaps in A's view of the common ground that do not arise in a conversation. This extra work is necessary just to understand the comment. To fully benefit from the comment by inferring the essay-writing principle M had in mind requires even more work.

The planned machine learning investigations will attempt to recognise and categorise appropriate opportunities for feedback comments by looking for associations between the categories assigned to each margin comment according to our scheme, and features of the passage in the essay to which a comment points—simple n-gram features, more complex measures of semantic similarity, and analysis of syntactic structure will be experimented with. The planned automatic feedback comment generator will be informed by the machine learning investigations. The form and style of the comments generated is yet to be decided.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (grant numbers EP/J005959/1 and EP/J005231/1).

References

Chris Anson. 1989. Response styles and ways of knowing. In *Writing and Response: Theory, Practice, Research*, pages 332–366. NCTE, Urbana, IL.

J.L. Austin. 1962. *How to do things with words*. Oxford University Press, Oxford, 2nd edition.

R.F. Bales. 1950. A set of categories for the analysis of small group interactions. *American Sociological Review*, 15(2):257–263.

E. Brown and C. Glover. 2006. Evaluating written feedback. In C. Bryan and K. Clegg, editors, *Innovative assessment in higher education*, pages 81–91. Routledge, Abingdon.

E. Brown, C. Glover, V. Stevens, and S. Freake. 2004. Evaluating the effectiveness of written feedback as an element of formative assessment in science. In C. Rust, editor, *Proceedings of the 12th Improving Student Learning Symposium*. The Oxford Centre for Staff and Learning Development, Oxford Brookes University, Oxford.

Harry C. Bunt. 1990. DIT: Dynamic interpretation in text and dialogue. In *ITK research report / Institute for Language Technology and Artificial Intelligence, 15*.

Harry Bunt. 2011. The semantics of dialogue acts. In *Proceedings of the 9th International Conference on Computational Semantics*. Oxford.

H.H. Clark and S.E. Haviland. 1974. Psychological processes in linguistic explanation. In D. Cohen, editor, *Explaining linguistic phenomena*. Hemisphere Publication Corporation, Washington.

H.H. Clark and E.F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, (13):259–294.

M. G. Core and J. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, MIT, Cambridge, MA.

M. Francine Danis. 1987. The voice in the margins: Paper-marking as conversation. *Freshman English News*, (15):18–20.

Dana R. Ferris, Susan Pezone, Cathy R. Tade, and Sharee Tinti. 1997. Teacher commentary on student writing: Descriptions and implications. *Journal of Second Language Writing*, 6(2):155–182.

Maxine Hairston. 1982. The winds of change: Thomas Kuhn and the revolution in the teaching of writing. *College Composition and Communication*, 33(1):76–88.

F. Hyland. 2001. Providing effective support: investigating feedback to distance learners. *Open Learning*, 16(3):233–247.

J.A.W. Kamp. 1981. A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stockhof, editors, *Formal methods in the study of language*, pages 177–321. Mathematical Centre Tracts, Amsterdam.

Wolfgang Klein. 1985. Ellipse, fokusgliederung und thematischer stand. In Reinhard Meyer-Hermann and Hannes Rieser, editors, *Ellipsen und fragmentarische Ausdrücke*, pages 1–24. Niemeyer, Tübingen.

D. Lewis. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, (9):339–359.

Erika Lindemann. 1987. *A Rhetoric for Writing Teachers*. Oxford UP, New York, 2nd edition edition.

Jason Merchant. 2004. Fragments and ellipsis. *Linguistics and philosophy*, (27):661–738.

Melissa M. Nelson and Christian D. Schunn. 2009. The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science*, 37:375–401.

Hadara Perpignan. 2003. Exploring the written feedback dialogue: a research, learning and teaching practice. *Language Teaching Research*, 7(2):259–278.

Allan Ramsay and Debora Field. 2008. Speech acts, epistemic planning and Grice's maxims. *Journal of Logic and Computation*, 18:431–457.

H. Sacks, E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, (50):696–735.

E.A. Schegloff. 1999. Discourse, pragmatics, conversation, analysis. *Discourse Studies*, 1(4):405–435.

J.R. Searle and D. Vanderveken. 1985. *Foundations of illocutionary logic*. Cambridge University Press, New York.

J.R. Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, Cambridge.

R Stalnaker. 1972. Pragmatics. In D. Davidson and G. Harman, editors, *Semantics of natural language (Synthese Library, Vol. 40)*, pages 380–397. D. Reidel, Dordrecht, Holland.

R. Straub. 1996. Teacher response as conversation: more than casual talk, an exploration. *Rhetoric Review*, 14(2):374–98.

R.H. Thomason. 1990. Accommodation, meaning, and implicature: Interdisciplinary foundations for pragmatics. In P.R. Cohen, J. Morgan, and M.E. Pollack, editors, *Intentions in communication*, pages 325–363. MIT, Cambridge, Massachusetts.

D. Whitelock, S. Watt, Y. Raw, and Moreale E. 2004. Analysing tutor feedback to students: first steps towards constructing an electronic monitoring system. *Association for Learning Technology Journal*, 11(3):31–42.

Nina Ziv. 1984. the effect of teacher comments on the writing of four college freshmen. In R. Beach and L. Bridwell, editors, *New Directions in Composition Research*, pages 362–380. Guilford, New York.