Listening. Learning. Leading.®

Research Report
ETS RR–12-16

Identifying Speech Acts in E-Mails: Toward Automated Scoring of the *TOEIC*® E-Mail Task

Rachele De Felice

Paul Deane

September 2012

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Identifying Speech Acts in E-Mails: Toward Automated Scoring of the *TOEIC*® E-Mail Task

Rachele De Felice and Paul Deane

ETS, Princeton, New Jersey

September 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Associate Editor: Joel Tetreault

Reviewers: Yoko Futagi and Jill Burstein

.

Abstract

This study proposes an approach to automatically score the *TOEIC*® Writing e-mail task. We focus on one component of the scoring rubric, which notes whether the test-takers have used particular speech acts such as requests, orders, or commitments. We developed a computational model for automated speech act identification and tested it on a corpus of TOEIC responses, achieving up to 79.28% accuracy. This model represents a positive first step toward the development of a more comprehensive scoring model. We also created a corpus of speech act-annotated native English workplace e-mails. Comparisons between these and the TOEIC data allow us to assess whether English learners are approximating native models and whether differences between native and non-native data can have negative consequences in the global workplace.

Key words: TOEIC Writing, e-mails, speech acts, pragmatics, workplace English

Acknowledgments

Thanks are due to Trina Duke, Liz Park, Waverely VanWinkle, Feng Yu, and Vincent Weng for enabling access to the TOEIC data; Peggy Auerbacher and Debi Quick for annotating the data; Russell Almond for assistance with statistical calculations; and Derrick Higgins for helpful comments in reviewing a draft of this report.

This study was conducted while Rachele De Felice was at ETS on an ETS Postdoctoral fellowship.

Table of Contents

List of Tables

Page

Introduction and Motivation

Good knowledge of e-mail text is a crucial skill for learners of English as a foreign language (EFL) to succeed in a global workplace because e-mail use lies at the heart of modern business communication. According to ETS (2010), the *TOEIC*® Writing test aims to prepare EFL learners to be good communicators in the workplace using a variety of writing tasks. In addition to photo description tasks and essay prompts, the test includes a dedicated e-mail writing task in recognition of the importance of this skill. The work discussed in this paper draws on a collection of TOEIC e-mail responses to introduce an approach to the automated scoring of this test item. We also present some observations on similarities and differences between native and non-native e-mails in the workplace domain to establish whether any differences may have unintended negative consequences for the non-native writer.

The assessment of EFL workplace e-mails involves several aspects of the text. For example, the TOEIC scoring rubric for this test item refers to the test-taker's awareness of the appropriate tone and register for the intended workplace domain, as well as his or her ability to carry out tasks such as asking questions, making a request, giving instructions, or conveying information. In the linguistics literature, these tasks are known as *speech acts* (Austin, 1962; Searle, 1969): the act of using language to perform a task. Speech acts are typically assigned to such general categories as statements, commitments, directives, and expressives (expression of a psychological state).

Research on EFL speech acts and pragmatics has been ongoing for many decades. Pragmatic competence is generally believed to be harder for learners to achieve than grammatical competence (see, e.g., Bardovi-Harlig & Dornyei, 1998; Kasper, 2001); at the same time, pragmatic errors can have more serious consequences than grammatical errors, since pragmatic errors often result in inadvertent impoliteness, which can harm communication more than a misused determiner or preposition. In recent years, researchers have begun to extend the analysis of pragmatic competence to new domains such as e-mail; as Andrew Cohen (2008) noted, "It is through e-mail that pragmatic failure is often found" (A. Cohen, 2008, p. 219). The present work is motivated by the importance of understanding pragmatic failure in e-mail communication, especially since few studies have been carried out in this field.

Our research also aims to make a contribution to the domain of natural language processing (NLP) and automated scoring. In particular, we present an approach to automatically

identifying speech acts in TOEIC Writing e-mails as a first step toward the automated scoring of these test items. The item rubric instructs the student to write an e-mail containing specific actions, such as asking for information or making a request. These items can be easily mapped onto basic speech act categories. One of the key criteria for scoring an answer is that all the speech acts required by the prompt should be present in the text; therefore, automatic speech act identification is a necessary component of automated scoring for this task.

To illustrate our point about the relationship between presence of speech acts and scoring, consider a set of responses where the rubric requires the response to contain at least two questions. Of these, 281 responses contain either one or no questions. The responses are scored on a scale of 0 (*lowest score*) to 4 (*highest score*). A score of 2 signifies a response with several weaknesses and 3 an unsuccessful or incomplete response. Of the 281 responses, 30% have received a score of 2 (accounting for 70% of all 2 scores in this set) and 27% a score of 3. While the lack of required speech acts is not on its own a sufficient criterion for the identification of lower-scoring responses, a strong link appears to exist between the two, suggesting that it may indeed be necessary to be able to recognize the presence or absence of speech acts for successful automated scoring.

Furthermore, a small but growing body of research shows speech act identification in e-mails using NLP tools, mainly for the purpose of analyzing and organizing incoming mail. Our work extends this line of research and applies it to non-native rather than native language. An ongoing debate questions the extent to which NLP tools and techniques can be successfully applied to learner language, so it is useful to examine whether this particular task is made more challenging by the presence of learner English (for a good introduction to these issues, cf. Meunier, 1998).

In particular, the NLP-based approach we propose relies on using a combination of resources including a parser, a part-of-speech (POS) tagger, and an n-gram list to extract syntactic and lexical information about each utterance. This approach is used to create one feature vector for each utterance. We train a machine learning classifier on the vectors to associate particular combinations of features to a given speech act category, allowing the classifier to automatically assign a speech act label to novel sentences. This method achieves over 79% accuracy on data taken from the same pool as the training material and up to 70% on data taken from a different corpus.

Although the main focus of the present work is a corpus and computational study of speech act data—and of the ways in which speech act data can best be represented and analyzed—one of its outcomes is a speech act–tagged corpus of learner e-mails that can be of use in research on second language acquisition (L2) speech act usage. This annotation allows more sophisticated analysis and filtering of the data. For example, information can be obtained about the usage patterns of different speech acts and whether learners appear to struggle more with some rather than others; or the naturalness and pragmatic acceptability of particular kinds of speech acts can be assessed to observe whether the required level of politeness or naturalness is being achieved by the learners (cf. De Felice, 2009). We can gain insights into the preferred phraseology of the learners in this domain to evaluate the breadth and appropriateness of their vocabulary.

Background and Related Work

Scoring and Pragmatic Errors

A rich body of work exists on automated scoring of essays and other kinds of test answers using NLP; a good introduction to the subject can be found in Deane (2006) or in the proceedings of recent dedicated workshops (e.g., Tetreault, Burstein, & De Felice, 2008; Tetreault, Burstein, & Leacock, 2009). Reasonable success has been achieved in detecting errors related to grammar and lexical choice (De Felice & Pulman, 2008; Gamon et al., 2008; Lee, 2009; Tetreault & Chodorow, 2008) and in assessing textual coherence in essay-length responses (Attali & Burstein, 2006; Burstein, Chodorow, & Leacock, 2003). Another area of research concerns scoring open-content short answers, which has proved more challenging because the amount of text may not always be sufficient to perform the appropriate statistical analyses. In open-content short-answer tasks, the focus of the scoring is different from essay tasks: The focus of assessment is on the accuracy of the content of the response rather than on the stylistic qualities of the writing, which requires an understanding of what constitutes correct or incorrect content. Furthermore, this content may be less predictable and therefore harder to compare to a set of predetermined model answers, as there may be more than one correct way of presenting the same information. Work at ETS (*c-rate*r™ scoring engine, Attali, Powers, Freedman, Harrison, & Obetz, 2008; Leacock & Chodorow, 2003; Sukkarieh & Bolge, 2008) and elsewhere (Pulman & Sukkarieh, 2005; Sukkarieh & Pulman, 2005) has attempted to address some of these challenges.

Nevertheless, testing is increasingly concerned with a variety of writing tasks typical of skills and activities that occur in everyday life, such as e-mails in a workplace context. In this domain, the text of the response is subject to some constraints: It must contain particular elements, relate to the topic, and exhibit formal correctness. At the same time, a relative degree of flexibility must be expected with regard to the content, since test-takers can introduce any number of facts in their writing to support their core message (for example, *I can't come to the meeting because I'm away/my child is ill/my car is broken* . . . and so on). Therefore, advance predictions of what kinds of sentences may or may not be correct in relation to the content are difficult. Furthermore, these tasks often aim to assess not only test-takers' grammatical proficiency, but also their pragmatic and communicative competence; that is, is the language they are using appropriate for the context of the task (e.g., not too informal or impolite). The new challenge for NLP applications in the learner domain, therefore, is how to deal with this kind of shorter, more open-ended text, including how to model its pragmatic features.[1]

It is important to stress, however, that pragmatic dimensions of foreign language learning and teaching are not a recent concern (a comprehensive overview of the field is beyond the scope of this paper; cf. for example Canale & Swain, 1980; Dornyei, 1995; Kaplan, 1966; Schmidt & Richards, 1979). Issues of relevance range from the acquisition and usage patterns of particular speech acts such as refusals, requests, apologies, or complaints (Biesenbach-Lucas, 2005; Kasper, 1984; Koike, 1989; Mey, 2004; Moon, 2002; Thomas, 1983) to more general studies on the perception of politeness and impoliteness in different groups of speakers. The latter is tied to the formulation of speech acts, whether in writing or in conversation (Mey, 2004; Thomas, 1983). Good recent overviews of the field can be found in Cohen (2008), Kasper and Rose (1999), and Kasper (Kasper, 2001).

Cohen (2008) emphasizes the importance of pragmatic assessment and of nurturing pragmatic competence in learners, especially with a view to their role as workers and communicators in a global workplace. Although Cohen recognized that "the means for assessing pragmatic ability are still in need of development" (p. 222), he believed that

> This should not deter teachers from including pragmatics in their instruction as it is an often high-stakes area for language learners where pragmatic failure in the L2 speech community can lead to frustrating situations such as completely misinterpreting what the boss wanted. (Cohen, 2008, p. 222)

Our work is a contribution to this growing field, showing a possible way of automatically assessing success in one of the areas of pragmatic competence, namely mastery of speech acts.

Speech Act Identification and NLP

As the field of speech act classification is still relatively young, a wide diversity is apparent in the data sources and the taxonomies used to identify speech acts. This diversity makes it difficult to directly compare results and establish what a state-of-the-art performance may be.

Khosravi and Wilks (1999) wrote one of the earliest papers on the topic, but they apply their method only on two small sets of 24 and 100 e-mails respectively. More recently, Lampert and colleagues (Lampert, Dale, & Paris, 2006, 2008a, 2008b; Lampert, Paris, & Dale, 2007) have investigated the issue of utterance-level speech act annotation in several papers; however, they do not go on to present results with e-mail data. Carvalho and colleagues (Carvalho & Cohen, 2005, 2006; W. Cohen, Carvalho, & Mitchell, 2004), on the other hand, do use e-mail data (researcher generated), but they annotate speech acts at the message rather than at the utterance level. Message- rather than utterance-level annotation is also preferred by Leuski (2005) and Goldstein and Sabin (2006). We chose to annotate at the utterance level for two reasons. First, for the purposes of answer scoring, we needed to be able to assign a speech act to each utterance, as the scoring guidelines refer to utterances rather than the message overall. Second, it was rare to find e-mail messages that fulfilled one function only (e.g., requesting or committing), making it difficult to select only one category to annotate an entire message.

Research by Corston-Oliver, Ringger, Gamon, and Campbell (2004) and Dredze and colleagues (Dredze, Brooks, et al., 2008; Dredze, Wallach, et al., 2008) gave practical examples of how this kind of work can be used in workplace settings. The former group proposed a system that can automatically identify the tasks in e-mails, based on speech acts, and turn them into to-do lists. The latter makes use of artificial intelligence and NLP techniques to develop tools that can assist e-mail users, such as reply and attachment predictors and keyword summarization. Mildinhall and Noyes (2008) followed a similar approach; they coded 144 e-mails from the Enron dataset (Berry, Browne, & Signer, 2007) according to speech act content in order to classify them as belonging to one of four e-mail genres: personal, business, admin, or inter-employee relations.

This brief overview makes it clear that considerable methodological variation exists on this topic, though all the papers mentioned take a similar approach to the task, namely use of a combination of features (including all or some of n-grams, POS tags, syntactic information, and punctuation) to create feature vectors as input to machine learning classifiers.[2] Our work compares to the current research in that it uses techniques similar to those described above. However, the goal of our research is to support answer scoring, rather than create a generic e-mail manager. Furthermore, we are concerned with non-native language, whose peculiarities, as noted above, can sometimes lead to impaired performance in NLP applications.

We have also collected a smaller corpus of L1 (native speaker) e-mails from the workplace domain. Although still in the early stages, this dataset enables us to compare more directly the structure of corresponding speech acts in native and non-native English. The L1 corpus serves as a useful benchmark to assess how well learners are approximating native models and to explore whether deviations from this model are detrimental to clear understanding and harmonious social relations.

The Data

The TOEIC E-Mail Task

According to ETS (2010), the TOEIC is an assessment that focuses on measuring English language skills for business:

> The TOEIC Speaking and Writing tests are designed in response to a need in the global workplace to be able to directly measure speaking and writing skills in English. Many global corporations and institutions need information about their employees' speaking and writing skills, as well as their listening and reading skills. (p. 4)

The TOEIC is, therefore, very different from other common tests of English as a foreign language such as the *TOEFL*® test or IELTS, in that TOEIC is specifically focused on workplace English. Currently, most research on L2 English relies on data obtained in academic settings, such as essays written by students in the course of their studies. The availability of data from a nonacademic domain offers the chance to view learner progress from a different perspective and, potentially, to highlight issues that would go unnoticed in more traditional learner corpora.

The present research focuses on one of the tasks of the TOEIC Writing test, which we refer to as *the e-mail task*. In this task, the test-taker is presented with an e-mail prompt referring

to such common workplace occurrences as meeting attendance, item ordering, or customer complaints. The directions instruct the test-taker to write an e-mail in reply, which must contain some combination of speech acts to achieve full credit. For example, the directions might indicate that the e-mail response must contain two requests and one offer of information, or two questions and one statement, or three requests. In scoring the response, we need to check for the presence of the speech acts indicated in the prompt; the system we describe can identify them automatically and thus is a key component of an automated scoring system for this test.

Our dataset is derived from two administrations of the TOEIC test—2007 and 2008. The e-mails in this dataset are a fairly homogenous set, as all were written by adult test-takers from East Asia (mainly Japan and Korea). However, test-takers did have a wide variety of educational and professional backgrounds. The majority achieved scores of 140–170 in the TOEIC Writing test (scores range from 0 to 200). The research presented in this paper is based on 1,163 texts, comprising over 125,356 words and 13,000 utterances.[3] Of these, we use 9,284 utterances, as the system we have developed does not currently consider the salutations and closings of e-mails, but only the content of the actual message. The average length of a text was around 12 lines. A typical answer looked like this:

Dear Davis,
I read your e-mail.
I would be free on Monday.
But I am just considering about several things.
How long time can I take your the regional office?
And can I use all of the facilities of your office?
It could be helpful for me.
Please, let me know as soon as possible.
Thank you.
Praternio

We can compare this to an L1 e-mail of a similar kind (from the TOEIC Reading test):

Dear Dr. Matsuoka,
Thank you for inviting me to present my research at your laboratory. Seeing you in San Francisco was a pleasure for me as well.
In regards to speaking in Fukuoka, I would like to speak on the earlier date if possible, since I have talks to give in Montreal and Toronto on September 27 and 28, respectively.
Given what I understand from our previous conversation, I will probably modify my usual presentation by spending less time explaining basic background information.
Instead, I will focus on the technical details of my research.
Please let me know what you think.
Thank you very much, and I look forward to seeing you again soon.

A more detailed comparison of L1 and L2 e-mails will follow below. However, we can already point to some evident differences between the two, which may affect the use of NLP tools. While bearing in mind that the L2 e-mail was produced under exam conditions and time constraints, we immediately observe that it is shorter and presents a much simpler syntax, with no subordinate clauses. The L1 e-mail, on the contrary, contains supporting details and information leading to more complex sentences. This difference can be an advantage for an NLP application focused on learner text as it makes its syntactic analysis less error-prone. Furthermore, the learner e-mails display relatively little variation in their choice of lexical and syntactic patterns. This, too, can be beneficial to NLP analysis as it means that the system can learn to rely on particular sequences as cues to a given speech act, for example. On the other hand, a system trained on learner data only may then not perform as well on L1 data, with its more complex syntax and richer lexical content. Our experiments included tests on two kinds of L1 data to establish whether this was the case for our system, too.

Annotation Scheme

As observed above, research in this field has not settled on a single annotation scheme. For example, some researchers focus only on requests and make distinctions within them (such as for information, permission, and action, as in Khosravi & Wilks, 1999), while others select categories typically associated with activities described in e-mails (such as requests, commitments, deliveries of files, as in Carvalho & Cohen, 2006; W. Cohen, et al., 2004). Others

choose to rely on a more general set of classes. Lampert et al. (2006), for example, initially based their work around the verbal response mode (VRM) framework (VRM, Stiles, 1992), originally designed to categorize psychology discourse. Rather than design a taxonomy from scratch, we, too, decide to use the VRM framework as our starting point because it suggests speech act categories that match linguistic classifications. The VRM scheme codes two aspects of speech acts: form and intent, roughly comparable to locutionary and illocutionary acts. These can coincide, as in a direct order—*Close the door!*—or not coincide, as in a very indirectly formulated order—*I wonder if it would be quieter if we closed the door*. These two aspects are represented by a pair of letters in each tag, where the first represents form and the second represents intent. While the surface form of an utterance is usually easily recognizable, it is often not clear what illocutionary act the speaker intended to perform. The relationship could be very tenuous between the two, where the illocutionary force of an utterance can only be correctly determined if knowledge of the wider context is available. For example, the utterance *I'm allergic to dogs* could be a simple statement of fact in the course of a conversation or a polite way of requesting that someone with a dog move away from the speaker. In the analysis of our system's results, we discuss the treatment of indirect speech acts to determine whether our taxonomy and the computational speech act model acquired by the system are able to treat these cases in the correct way.

In Stiles's original taxonomy, many factors assist in the identification of an act, including the grammatical/surface form of an utterance. We retain some of the key elements of the scheme, but we make some modifications and simplifications to adapt the system to e-mail data.

Table 1 shows the six categories used. While the brief descriptions given for each class are mainly centered on their content and function, formal aspects, such as syntactic and lexical characteristics, are also considered. A full description is given in Appendix A.

Annotation Procedure

Three coders were involved in the annotation procedure: one of the authors, and two others who were familiar with the test material but not with this particular coding scheme. They were issued detailed coding guidelines prepared by one of the authors. Each e-mail was divided up into utterances, with each utterance receiving one code only. The first author coded all the data, while the other two coders received a subset of 805 e-mails each, including 447 overlapping ones for the purpose of calculating inter-annotator agreement.

Table 1

Categories Used for Speech Act Annotation

Class	Brief definition	Example
AA – Advisement	Requires action from the hearer or change in his or her mental state	Request: *Please send me the files.* Imperative: *Don't worry about the meeting.*
DD – Disclosure	1st person statement sharing thoughts of speaker – cannot be verified empirically	*I am so happy to see you* *I have a question for you*
DE – Commitments/ factual statements	1st person statement, conveys fact that can be verified empirically	*I will attend the meeting* *I was there last week*
QA – Advisement in question form	Usually a polite or indirect request/order; can be embedded	*Could you send me the files?* *I was wondering if you could call me.*
QQ – Simple question	Genuine request for info – requires no action on the part of the hearer	*What's your name?* *I want to know what time it is.*
OT – Other statements	Exclamations, 3rd person statements	*The meeting is tomorrow.* *You must be so busy!*

Many discussions of annotator disagreement can be found in the literature, but this topic is too extensive to be discussed in the scope of this report (a comprehensive survey is Artstein & Poesio, 2008). Annotator disagreement is even more likely to surface in a pragmatic interpretation task such as this one, which is often open to multiple interpretations, even with sufficient context. For example, the utterance, *Do you know how to use this?*, could be a genuine inquiry, a request for help, an indirect way of asking the hearer to use the object in question, or even a subtle way of encouraging the hearer to stop using something.

Therefore, even with appropriate training materials, we cannot expect very high levels of agreement among annotators. Indeed, on a set of 1,606 utterances annotated by one of the authors (R1) and the two annotators, agreement as measured by Cohen's kappa (J. Cohen, 1960) was as follows:

R1 – R2: 0.673

R1 – R3: 0.693

R2 – R3: 0.638

Average: 0.668

These figures are considered good, as are those found in other work: Mildinhall and Noyes (2008) reported a kappa of 0.76 and Cohen et al. (2004) had an average of 0.77, as did Lampert et al. (2008a), although their annotators had a choice of two classes only. All observe variations in agreement depending on the class. For example, both Lampert et al. and Cohen et al. reported a much higher agreement for requests than for commitments, a trend that is also clearly observed in our data. In the discussion of the classifier results, we will determine whether these confusion patterns are typical just of human annotators or of automated ones, too.

We compared the three annotators' judgments to create the final version of the tagged data. For the instances that received three annotations, the majority tag was chosen where present; in case of disagreement, and for those that only received two annotations, the choice defaulted to that of the more experienced annotator. In the data thus obtained, there is some disparity in the size of the various classes, as can be seen in Table 2, which reports figures for the training data.

Table 2

Distribution of the Six Classes in the Training Data

Class	Number of instances	% of total
DE	2,429	32.71%
DD	1,270	17.10%
QQ	1,204	16.22%
AA	1,140	15.35%
OT	859	11.57%
QA	523	7.04%
Total	7,425	

Note. AA = advisement, DD = disclosure, DE = commitment/factual statement, OT = other, QA = advisement in question form, QQ = simple question.

Methodology

The Feature Set

The task of identifying features that can be considered characteristic of various speech acts can be approached in different ways. In their treatment of the subject, standard pragmatics textbooks (Cruse, 2000; Jaszczolt, 2002; Levinson, 1983) refer to several elements of the utterance, such as sentence type (declarative vs. imperative vs. interrogative), lexical items such as *please* or certain kinds of verbs (e.g., *order*, *request*, *promise*, *declare*), verb mood and tense. However, these are not considered by the authors to be exhaustive criteria for speech act identification,[4] nor are they intended to be of direct use for NLP research, and so they are not on their own sufficient.

As observed previously, NLP research in this area has broadly converged on a set of characteristics found to be predictive of speech act categories, including punctuation, lexical items, n-gram sequences, and syntactic structure. In fact, many of these are also mentioned in the guidelines for our annotators (e.g., noting whether there is a first or third person subject, or whether there are any modal verbs). Therefore, for the purpose of designing a set of features to be used in a machine learning classifier, it seems appropriate to take into account the insights from all the sources discussed above: textbooks, related literature, and annotation guidelines.

In addition, to avoid bias in the dataset, which might occur if one were only to rely on preconceptions of what should be useful for the task, we also manually inspected a large sample of instances from each class after annotation. If a given characteristic appeared to occur repeatedly, it was also captured as a feature. For example, we noticed that imperatives (AAs) often begin with the word *please* and that question-requests (QAs) often begin with the verb *can* or *could*. These characteristics suggested that important regularities could be accounted for by including a feature that noted the first word of the utterance.

The system we are developing is intended to be fully automated; it is therefore crucial that we are able to extract the relevant information about the utterance automatically, too. Indeed, all the elements mentioned above are easily recoverable using a suite of NLP tools. The C&C toolkit[5] (Curran, Clark, & Bos, 2007) consists of a set of applications including a morphological analyzer (morpha, Minnen, Carroll, & Pearce, 2001), a POS tagger, a parser (Clark & Curran, 2007), and a named entity recognizer (NER, Curran & Clark, 2003). Together, these tools allow us to record information about the tense of verbs in the sentence, distinguish

modals from main verbs and imperative, declarative, and interrogative sentences from each other, and note the presence of proper nouns, among other things. Furthermore, the parser's output also includes the grammatical relations among the various components of the sentence (e.g., subjects and verbs, verbs and objects). These relations are very important as they allow us to correctly identify the subject, predicate, and object of the utterance. This information is not always recoverable by just considering word order, and it also forms part of the feature set, as we will explain below.

Overall, 19 different kinds of features are considered as summarized in Table 3. While some of these features can only have a fixed number of values (e.g., the possible POS tags of verbs are a finite set), the value of many others can be represented by any lexical item. Each feature will be described briefly in turn, followed by some detailed examples.

Feature Description

Subject type and subject item. The possible value of none for subject type is included to account for imperative sentences. It is evident that the kind of subject is a key distinction among the different classes. Noting what kind of pronoun is the subject serves to further distinguish between first and third person categories.

Object type and object item. These features are included because arguably categories involving actions are more likely to contain transitive verbs with direct objects than those involving feelings or thoughts: cf. *I will send **the papers**, Please call **me** tomorrow* vs. *I am happy about the decision, I feel ready for the meeting*.

Presence of modal and modal item. Requests and commitments very frequently include modal verbs (cf. ***Can** you call me?, I **will** write the report tomorrow*).

First and last word. Manual inspection of the data revealed that many instances within the same class tend to begin with the same word, prompting us to include the first word of the utterance as a feature. Additionally, there may also be shared similarities with regard to the final word of the utterance; for example, temporal adverbs tend to be placed in a sentence's final position (cf. *I will be there **tomorrow**, I am going to the meeting **tomorrow***).

Verb type and verb tag. The verb plays a key role within a speech act, with different kinds of verbs determining the intention of the utterance. Imperatives will be found in request and orders; for example, to-infinitives may be found often with reference to actions to be undertaken (*I am going **to attend** the meeting*), while bare infinitives are typical in QAs (*Can*

Table 3

Features Used for Vector Creation

Feature	Possible values
Punctuation	; , . ! ? none
Length	Length of utterance
Subject type	Type: noun, pronoun, none
Subject item	Item: if pronoun, which one
Object type	As above
Object item	
Has modal	Yes/no
Modal is	Can, will, would, should, etc.
First word	Lexical item (excluding punctuation)
Last word	
Verb type	Infinitive, participle, etc.
Verb tag	VBD, VB, etc.
Sentence type	Declarative, question, embedded, etc.
Has wh-word	Who, what, when, why, where, how
Predicative adjective	Yes/no
Adjective is	Lexical item
Complex structures	I + modal + inf
	Please + imperative
	Be + adj + for me
	Etc.
Named entities	Place, time, name, organization, date, money
Unigrams	Small set of distinctive n-grams – cf. text
Bigrams	

*you **tell** me his number?*). The NLP tools we use, described below, allow us to capture this syntactic information (which we can refer to as the mood of the verb), labeling each verb as a to-infinitive, passive or active past participle, gerund, or bare infinitive/imperative. In addition, it is also important to record more basic information about the verb, such as its person and tense; the POS tags for the verbs give tense information and, for the present tense, also distinguish between third person and non-third person. The usefulness of tense is justified by the fact that certain tenses are more likely to occur with particular categories; for example, the past tense, referring to events that have already occurred, is more likely to be found in first or third person statements

referring to facts (commitments/factual statements [DEs], other statements [OT]), while the present participle is frequent in commitments that do not have modals (*I am **coming** to the meeting*).

Sentence type. Similar to the verb phrase categories, our tools also assign nine sentential categories: declarative, wh-question, question, embedded question, embedded declarative (*He said **he is coming***), fragment, for-clause (***For him to say so**, it must be true*), interjection, and elliptical inversion (*I was at the meeting, **as was Mary***). Having this kind of information available can be very useful, especially as an additional tool to distinguish simple questions (QQs) and QAs from the other categories.

Wh-words. Found in direct and embedded questions.

Predicative adjectives. Adjectives following a verb such as *be* or *feel* tend to be more frequent in disclosures (DDs), which refer to feelings and thoughts, than in more action-based speech acts: cf. *I **am sorry** to miss the meeting, I **feel ready** for the challenge*. Information about the actual lexical item present is also included, because while some adjectives, such as the ones above, are more likely to refer to human subjects, others, such as *impossible* or *urgent*, are more likely to refer to inanimate subjects, which would belong to a different speech act category (e.g., *It **is impossible** to have a meeting next week*).

Special or complex structures. This shorthand term for a feature aims to exploit the fact that speech acts are formulaic: We often find particular combinations of syntactic and lexical elements used to realize a given speech act. This observation is borne out not just by analysis of the data at hand, but also by discussions in the literature, both in general linguistics textbooks and in those aimed more specifically at learners of English. These combinations often offer pointers to help identify or formulate particular acts; for example, they might say that requests tend to include the word *please* and a modal verb (cf. ***Can** you **please** call me?*). They can be considered the basic way of representing many speech acts, and in most cases the level of linguistic sophistication of the test-takers is not so high that they deviate much from these structures. It is possible that L1 data does not rely on these combinations so heavily.

Our feature set includes six such structures:

1. I + modal + (bare) infinitive (*I can come*, *I will do this*), typical of commitments
2. Modal + you + (bare) infinitive (*Can you come?*, *Will you do this?*), typical of advisement questions

3. Be + adjective + for me (*It is good for me*, *It is important for me to be there*), typical of a subset of the general other (OT) category
4. I + adjective + (to) infinitive (*I am happy to help*, *I will be honored to attend*), also typical of commitments
5. Please + imperative (*Please call me*), typical of advisements
6. Am + present participle (*I am going to be there*), typical of commitments, as suggested above

Some of these categories represent information that is already partially captured by other features. For example, both the presence of modals and of particular verb forms is already recorded in separate features. However, consolidating this information into a single feature representing syntactic structures typical of a particular class may be more predictive than having the individual pieces of information alone: This feature highlights the fact that the various elements belong to the same phrase and are therefore a meaningful unit, rather than just being present in the sentence but otherwise unrelated.

Named entities. Some named entities, especially times and dates, can be often found in speech acts involving a commitment (either on the part of the speaker or the hearer), as in *Please come to my office on* ***Monday*** or *I will attend the session at* ***2 pm***.

N-grams. All automated approaches to this task discussed above include n-grams among the features used for classification. The repetition of phrases and key words typical of this genre and this task is such that the use of n-grams lends itself particularly well to the identification of the various categories. We focus on unigrams and bigrams, excluding longer sequences, to avoid issues of data sparseness.

Although our dataset contains 116,939 tokens, its type/token ratio is extremely low: We have only 3,209 distinct types, giving a type/token ratio (equal to (types/tokens)*100) of 2.74[6] for the whole corpus; the token/type ratio is 36.44. These figures suggest that the vocabulary of the responses is not very diverse, with a high degree of reuse of many lexical items. However, these figures are not necessarily representative of L2 e-mail writing as a whole: It must be remembered that these texts are very homogeneous, having all been written in response to a single test prompt; a corpus containing e-mails on a diversity of topics would be more likely to include greater variation. For example, preliminary analysis of a subset of the Enron e-mails

suggests that native e-mail text may indeed be richer and more complex: In 250 e-mails we find an average of 991 tokens and 324 types per e-mail (type/token ratio 32.69). Furthermore, these measures are very sensitive to length and tend to be less informative when texts are very short, as is the case for the e-mail responses. Although direct comparisons with other sources are hard to come by, given the scarcity of datasets with a level of homogeneity comparable to ours, the figures point to the conclusion that L2 e-mails, at least in this domain, are lexically poor. They also reinforce the claim that a high degree of reuse of phrases and repetitiveness is present in this type of text. Therefore, it is reasonable to assume that the presence of particular lexical items is a strong indicator of class.

Rather than including all available unigrams and bigrams in the feature set, only the most informative ones are selected (Carvalho & Cohen, 2006; Khosravi & Wilks, 1999; Lampert, et al., 2006). This selection prevents the feature space from becoming too large (and therefore allowing faster performance) and avoids the inclusion of data that are likely to be irrelevant to the task. Furthermore, as we are interested in developing an approach that can be generalized to any number of prompts, it is important to discard any n-grams in our corpus that are prompt-specific. For example, if the prompt required the answer to discuss water bottles, the n-grams *water, bottle,* and *water bottle* would be discarded.

The goal is to obtain a set of words that are less common (and less likely to be found in almost any text) and most representative of a single speech act. To achieve this goal, in constructing the unigram set, we selected only those types that did not occur in every class—as these were more likely to be typical of a single speech act only—and calculated the tf-idf scores for each type[7] (cf. Manning & Schuetze, 1999, p. 541ff.). Tf-idf scores are a useful discriminating element as they are higher for those terms that are found in one subset of documents only. If we think of our dataset as containing several subsets, each of which represents one speech act category, it follows that a good criterion for selection is to focus only on words that have a tf-idf score above a certain threshold for a given class. We applied this criterion as described below.

For each speech act class outlined above, we considered the 50 unigrams (words) with the highest tf-idf scores, and of those, we selected for inclusion only the ones that either belonged to one class only, or if they belonged to more than one class, did so with much lower frequency in one than in the other (by a factor of 10 or more). For example, *instruction* appears

in the word lists for both the AA (direct request or order) and QA (request-as-question) classes, but while the score for the AA class is only 0.000062, for the QA class it is 0.000241 (27th highest for that class), leading to its inclusion in the set of QA-specific unigrams.

The final set of 84 unigrams, and the classes they are associated with, is reported in Appendix B. The words included are not particularly rare or unusual, but upon reflection, they are typical of particular speech acts (cf. for example, *promise*, *willing* for commitments, or *happy* and *wish*, for expression of feelings). The unigram data can be used in two features: either simply noting the presence of a term belonging to a particular class or the presence of the term itself. It is possible that the approach used has been too conservative and that retaining a larger number of unigrams, or selecting them differently, would have lead to better results; future experiments will be needed to assess this.[8]

The selection of bigrams followed a somewhat different procedure that did not make use of tf-idf scores. Stopwords were not removed, as these include items such as modal verbs, auxiliaries, and pronouns that, as we have mentioned, play an important role in class identification; however, in further developments of this tool we will remove determiners and perhaps prepositions. Again, after discarding prompt-specific bigrams, for each class we considered the remaining bigrams with a relative frequency of at least 0.01. We compared these sets to identify bigrams that occurred in one class only and those that occurred at most in one additional class with similar frequency.

Then, to ensure that these bigrams are typical of workplace e-mail, rather than e-mail or L2 writing in general, two further steps were carried out. First, we made use of the Enron e-mail corpus (Berry et al., 2007).[9] We identified and annotated a subset of e-mails from that corpus whose content was comparable to that our own data, namely discussions of everyday business. Due to practical limitations, this corpus was not very large, consisting of only 250 e-mails. We extracted bigram counts for each class from the Enron data and compared them to the bigram sets for each class of the L2 e-mail data. If a bigram was found with high frequency for a given class in both Enron and TOEIC data, it was retained. As mentioned above, the goal is to obtain a set of bigrams that are characteristic of e-mail workplace discourse, of which both the Enron e-mails and our own data are examples. Therefore, presence in both datasets offers a stronger guarantee that the bigrams belonging to this discourse.

However, as a further check, their frequency in the e-mail domain was compared to that from the written part of the American National Corpus (ANC, Reppen, Ide, & Suderman, 2005),[10] which can be considered representative of a more general domain of American English. If a bigram is truly typical of workplace e-mail discourse rather than language in general, its frequency in the e-mail corpora should be much higher than its frequency in the ANC.[11] The e-mail bigrams were compared against a list of the top 500 bigrams in the ANC: Any that appeared in both sets were discarded. After submitting the data to these analyses to ensure their domain-specificity, we were left with a set of 17 bigrams only. These are shown in Table 4, together with the class(es) they belong to. Again, it is possible that this approach may prove to be too conservative and that potentially useful information might have been lost in this way.

Table 4

Bigrams Used

Bigram	Class	Bigram	Class
I will	DE	could you	QA
I hope	DD	can you	QA
I think	DD, DE	do you	QQ, QA
I am	DD, DE	is it	QQ
I can	DD, DE		
let me	AA, QA	it is	OT
me know	AA, QA	there is, are	OT
please let	AA		
call me	AA		
you can	AA		

Note. AA = advisement, DD = disclosure, DE = commitment/factual statement, OT = other, QA = advisement in question form, QQ = simple question.

All these sequences contain at least one stopword and yet have been identified as the most representative for the various classes, which supports the decision not to remove stopwords for this procedure. Despite being very simple phrases of the language, their class-specificity is clear; indeed, these could almost be considered the basic building blocks of a speech act, which assist in clearly signposting the speaker's intention.

Examples

Two sample sentences follow to illustrate how the features described above relate to the various elements of the sentence itself. For the purpose of clarity, not all features are given for each sentence.

1. I am so happy about this. [tag: DD]

   Punctuation: .
   Length: 7
   Subject: pronoun, *I*
   Sentence type: declarative
   Predicative adjective: is present, *happy*
   Verb tag: VBP (non–third person present)
   First word: *I*
   Last word: *this*
   Unigram: *happy*

2. Of course I can help you. [tag: DE]

   Object: pronoun, *you*
   Modal: yes, *can*
   Verb type: bare infinitive, *help*
   Complex structure: I + modal + infinitive
   Bigrams: *I can*

Feature Vector Creation

To create the feature vectors, we relied on the tools introduced above. Each utterance was run through all the C&C tools described above; a Python script then processed the output to extract the relevant information, populating the feature vector for each utterance by relying on the information supplied by the tags, the grammatical relations, and the lexical items themselves.

Our full-featured training set has 5,885 different feature-value pairs. This is the pairing of a type of feature with its value in that particular context, such as "`'FirstWord': 'I'`." A typical feature vector is shown below in Figure 1.

Experiments and Results

Classifier Choice

All the experiments in this study were carried out with a maximum entropy classifier (Ratnaparkhi, 1998); we used the implementation found in the Natural Language Toolkit (Bird & Loper, 2004). Maximum entropy algorithms are often used in NLP because they do not make any assumptions about feature independence. This is important given the design of our feature set, where it is clear that some features are interdependent; for example, if the feature "has bigram: I can" is present, it follows that the feature "has modal: can" will also be present.

```
I fully understood your situation [DD]

      [{'Modal': 'no', 'LastWord': 'situation', 'Object':
      'noun', 'HasUnigram_understood': 'yes', 'Punct': 'none',
      'Length': 5, 'PredicativeAdj': 'none', 'VerbTag': 'VBD',
      'FirstWord': 'I', 'SubjectIs': 'I', 'SentenceType':
      'S[dcl]', 'Subject': 'pronoun'},'DD']
```

Figure 1. **A typical feature vector.**

The basic experiment design is as follows: The classifier was trained on a set of about 8,000 feature vectors (training set), each belonging to one of the six possible classes; it was then tested on a held-out set of about 800 vectors, to which a class must be assigned (test set). Its performance was measured in the first instance in terms of *accuracy*, which refers to the percentage of times the correct class was assigned to an instance. We also measured the system's *precision* and *recall*, which we can express in terms of true and false positives and true and false negatives (cf. Manning & Schuetze, 1999 ch. 8). A true positive (TP) is a correctly classified instance. A false positive (FP) for a given class is an instance that has been incorrectly labeled as belonging to that class; for example, a DD instance classified as DE is an FP for DE. Similarly, a false negative (FN) for a given class is a *missed instance*—one that should have been labeled as belonging to that class but wasn't. So in the example of the previous sentence, that instance would also count as a FN for DD.[12]

Precision measured, for a given class, how many of the instances classified as belonging to that class actually did belong to that class; in other terms, how much *noise* had been introduced. It is equal to TP/(TP+FP). *Recall* measured how many of the instances belonging to any given class had actually been labeled as belonging to it; in other words, how much the classifier has missed. It is equal to TP/(TP+FN). We also report the *f-measure*, which is the harmonic mean of precision and recall; its formula is (2 * Precision * Recall)/(Precision + Recall).

Precision, recall, and f-measure are standard performance measures for many NLP tasks, including related work in the area of speech act identification, making it easier to compare the success of our application against those of others. Their frequent use is due to the fact that they provide a more realistic measure of the application's success, giving information not just about the things that it is getting right, but also quantifying its errors in the form of FPs and FNs. For example, our application could achieve 99% accuracy in identifying requests, but it could be doing so by simply labeling every instance as a request, thus accumulating a large number of incorrectly labeled instances in the process. By reporting accuracy only, this flaw in the system would be obscured: precision and recall, on the other hand, would allow this information to be known. In this particular example, precision would be very low.

Weighted kappa (J. Cohen, 1968) is another measure often used in quantifying agreement between two judges, whether humans or computers. It is especially useful in contexts where the judges are required to make a score prediction along a scale of values, as it takes into account the distance along a scale of two different judgments and not just the fact that they are different per se. However, since the speech act identification task only calls for binary yes/no judgments, we do not feel it was necessary to include this measure as part of our evaluation metrics, though we report simple kappa.

Establishing a Baseline

Different studies have relied on different combinations of features to establish a baseline for the task, such as one-level decision trees (Lampert et al., 2006) or weighted bag-of-words (BOW) models (Cohen et al., 2004).[13] The most basic approaches, such as random guessing or selecting the majority class, would give very low results. Random guessing would only be successful one in six times, while selecting the majority class, DE, would have a success rate of 32.2%.

A more reasonable baseline might be one that helps us understand how much we gain by introducing all the extra layers of NLP and feature extraction, to see whether the increase in processing requirements is justified by an increase in performance. In this respect, such a baseline might be one that uses very basic information such as only punctuation, or the first word, or the selected bigrams and unigrams (which we recall are not a very large set): in other words, no syntactic processing. Accuracy scores for these runs are shown in Table 5. The results are averages of 10-fold cross-validation: The data is divided into 10 subsets, and each subset is held out in turn while training is performed on the remaining 9. This averaging is done to obtain a more representative estimate of the system's error rate.

Additionally, as in related work, we also implement an unweighted BOW baseline. The BOW feature vectors record, for each instance, the presence or absence of the 400 lexical items in the corpus with a frequency of 10 or more (we exclude determiners but not other stop words), without any further information as to their frequency, ordering, or the structural relations between each other.

Table 5

Baseline Results

Features	Accuracy
Punct	45.30%
firstword	53.87%
lastword	43.09%
firstandlast	59.62%
n-grams	49.75%
ngrams+firstw	67.57%
BOW	75.10%

Note. BOW = bag of words.

A significant difference exists (paired *t* test, $p < 0.001$) between all these scores; the only nonsignificant difference is between the punctuation only set and the last word only set. As shown in Table 5, a single kind of feature on its own is clearly not sufficient to guarantee an acceptable performance. The highest baseline score is given by the combination of two kinds of lexical features, the first word and the n-grams. The significant difference between this baseline and the others points to a key fact: A central role in distinguishing speech acts is played by the lexicon, and indeed arguably a rather restricted set of lexical items.[14] The simple BOW baseline

appears to be higher performing than all the other baselines, reinforcing the notion that the lexicon plays a key role in this task. However, as we will see in a later section, a model trained on lexical items only may be too closely linked to the training data and may therefore not generalize as well to novel data.

Improving on the Baseline: The Other Features

In this section, we describe the performance gains obtained by introducing additional features to the n-gram + first word baseline of 67.57%. An overview of the results is in Table 6, which gives the averages over 10-fold cross-validation. Our best result, using all available features, is 79.28% accuracy.

Table 6

Performance Gains With Additional Features

Additional features	Accuracy
A: Combined baselines	70.66%
B: A + subject	73.98%
C: B + last word	76.31%
D: C + object	76.91%
E: B + length	76.33%
F: B + modal	76.97%
G: F + verb info	77.85%
H: G + sentence type	77.56%
I: H + object	78.30%
J: I + complex struc.	78.49%
K: J + adjectives	78.68%
L: K + NER	79.18%
M: L + wh-words	79.36%
All (incl. length)	79.28%
Bag of words	75.10%

The first obvious addition is combining the features used in the baselines, namely n-grams, first word, and punctuation. This combination brings an improvement of 3% over the baseline (accuracy 70.66%), which is only significant for $p < 0.005$ (here and throughout this discussion—paired *t* test; $p = 0.004$). However, a real difference is noticed with the introduction of syntax-related features, such as the subject information: Its addition improves accuracy over the baseline by almost 6% (a significant improvement, $p < 0.001$). This improvement is not

surprising, since the speech act categories are clearly distinguished on the basis of who is performing or supposed to be performing the action. The linguistic intuition is confirmed by the results, and the gain in performance supports the decision to include parser output in the feature set. However, further syntactic information such as that relating to objects is less important and does not bring a significant improvement.

The role of the last-word feature is less clear. Its low performance when used on its own suggests that this feature is not very useful, but when used in conjunction with other features, it appears that it can play a valuable role, as demonstrated by the 3% gain in accuracy obtained over the plus subject set. However, this improvement is only significant for $p < 0.005$ ($p = 0.002$). Surprisingly, length does not appear to be an informative feature for this task. Indeed, if we compare the accuracy figures for the all set and the M set (which has all features except for length), we find a marginally higher score for the set without the length feature. We expected that certain types of speech acts would be shorter than others (e.g., statements of facts could be very verbose, while QAs could be shorter). It is possible that the tendency of our pool of test-takers to prefer simple syntax in their writing has led to all utterances being a similar length. We exclude length as a feature for the remaining experiments described in this section.

Given our earlier considerations on the distinctive use of modal verbs in speech acts, especially commitments and requests, it seemed reasonable to expect that their addition to the feature set would bring a significant improvement in accuracy. In fact, we find that adding them to the plus subject set (without the last word feature) increases accuracy by only 3% ($p = 0.002$). We attribute this increase to the large amount of information about the presence of these words that was already captured by the n-gram features, which included modals among the lexical items recorded; an additional dedicated feature does not contribute much extra information. The other verb features seem even less informative. These include both verb tags (person/tense) and category tags (essentially indicators of mood). The 1% improvement associated with this feature is not significant, which suggests that though differences in verb choice are an important characteristic of speech acts, they are not the main distinguishing one.

All the other features bring only marginal incremental improvements to accuracy, none of which are significant. This performance may be due to the fact that they are merely reduplicating information already present in other features, although the linguistic intuitions underlying their inclusion remain valid. However, it is important to point out that although these features in

isolation are not determining correct classifications, taken together they do allow higher accuracy. In particular, the full dataset accuracy (79.28%) is significantly better than the baseline + punctuation, plus subject, and plus last word sets (all $p < 0.001$), showing the value of including a variety of linguistic features. Although not all features may be equally important for all classes, together they all contribute to more accurate speech act identification.

We also note that there is not a very large performance gain between using all features and the BOW baseline. This finding is not entirely surprising, as many of the features are based on lexical items (e.g., n-grams, modals, first and last words), so some overlap occurs in the information captured by the two models. However, the top 10 features by information gain (listed below) for the full-featured set suggest that there is also some value in the way this set represents the lexical and structural properties of the utterances as meaningfully interrelated units. It contains several syntax-based features, showing that syntactic information does play a role in distinguishing among speech act classes and offering support for the linguistic intuitions underlying our feature choice. This information can be crucial when training and testing data differ extensively in content, such that lexical similarities alone are not sufficient to guarantee good performance; this hypothesis would require testing on a dataset such as personal e-mails, which is difficult to acquire.

The top 10 features by information gain are as follows:[15]

1. FirstWord
2. SentenceType
3. LastWord
4. ModalIs
5. VerbTag
6. Punctuation
7. SyntacticStructures
8. WhWord
9. SubjectIs
10. HasUnigram please

Simple kappa between the classifier and human annotations is 0.784, which is considered good. We observe that it is somewhat higher than the average inter-rater agreement of 0.668, which may be due to the fact that, unlike humans, the classifier is less likely to consider several different possible interpretations for the same utterance and be influenced by context or other external factors. Good agreement between classifier and human raters gives further support to the claim that this system can play a role in automated scoring.

This section presented a brief overview of the contribution of each feature to overall accuracy. The results indicate that lexical information in the form of n-grams and first/last words together with subject information are key indicators of speech act classes. However, the extraction of further features through syntactic processing of the data enables us to obtain significantly higher accuracy, justifying our decision to include these linguistically motivated features in our set.

Result Breakdown by Class

Table 7 gives the results for individual classes when using the full feature set. The figures for this table are taken from one of the folds to illustrate the major trends occurring in the data. Similar performance across runs suggests that the patterns found here are typical of the dataset as a whole.

Table 7

Precision and Recall for Individual Classes

	Right	Wrong	Total	Recall	Precision	F-score
Imperatives (AA)	123	20	143	86.01%	90.44%	88.17%
Expression of feeling (DD)	118	44	162	72.84%	71.52%	72.17%
Commitment or verifiable statement (DE)	253	44	297	85.19%	82.41%	83.77%
Third person statement (OT)	94	13	107	87.85%	80.34%	83.93%
Request as question (QA)	45	19	64	70.31%	95.74%	81.08%
Open question (QQ)	131	18	149	87.92%	87.33%	87.63%
Total	764	158	922			
Average				81.69%	84.63%	82.79%

Note. AA = advisement, DD = disclosure, DE = commitment/factual statement, OT = other, QA = advisement in question form, QQ = simple question.

We can see from Table 7 that average precision (macro-average, 84.63%) and average recall (macro-average, 81.69%) differ only by 3%. Often in tasks related to assessment, precision is favored over recall, as the effects of not catching some errors are considered to be less severe than those of giving erroneous feedback (e.g., identifying an error when none exist). It appears here that we are achieving a good trade-off between the two measures, though it is possible to further increase precision and observe the effect it has on recall.

We also observe that the individual F-scores range from just over 81% (QA) to over 88% (AA), with the exception of DD, which presents a rather lower score of 72.2%. Differences in performance across classes are not surprising; other work in the area reports similar divergences, and indeed even human annotators, as discussed above, often disagree about the nature of particular categories. Therefore, achieving F-scores above 80% for almost all categories is an indicator of good performance for the system at this stage of development. At the same time, it is important to understand the causes for these divergences: Are some classes more inherently difficult for the classifier to acquire models for, for example? Or are underlying flaws present in the system's development? In the next section, a detailed error analysis attempts to address these issues.

As a general observation, we note that requests—both in the form of imperatives and questions—are the most clearly identifiable (cf., the highest precision), though some variation must exist within them, which causes a lower recall. Conversely, the lower precision scores for commitments and third person statements suggest that some noise exists in the data that is misleading the classifier. Similar patterns were also identified by related work. The DD class is clearly the hardest to identify correctly, while QQs seem to be more immune to precision and recall issues. It must also be noted that the number of instances of QAs is very low, which could distort the figures somewhat.

The error analysis in the next section might also shed light on whether a correlation is apparent between the amount of data seen in training and the results. As we saw in Table 2, a certain amount of disparity is present in the number of instances available for each class in the training data, from just over 500 for QA to over 2400 for DE. However, with the exception of QA, where the number of instances is small and performance is not very high (suggesting an insufficient amount of data for training purposes), it is not clear that such a correlation between

number of instances and performance exists. For example, DD is the second largest class, but has the lowest F-score.

How Does This Compare?

Because of the differences in approaches and datasets found in the field, it is difficult to make direct comparisons among the various systems developed. These differences regard not only the data used in training and development, but also the way the systems are tested and how their performance is assessed. The use of different taxonomies also prevents us from directly comparing our results with other systems for individual classes to see whether our findings are consistent with the results presented literature about other systems.

Other systems, for example, Khosravi and Wilks (1999) reported accuracy of 90.28% for e-mails from the same domain as the training data and 79% for out of domain data, but these results referred to only 24 and 100 e-mails, respectively. Lampert et al. (2006) used various combinations of classifiers and feature sets, and the highest result they obtain is 79.75%. They also reported precision and recall figures for all eight classes, which average 70.65% and 71.76% respectively (average F-score: 70.82%). Their test set consists of just over 1,300 instances, which would make it more closely comparable to our own; but, as their data comes from discourse rather than e-mail data, as outlined above, a direct comparison is also not possible. Carvalho and Cohen (Carvalho & Cohen, 2006) and Cohen et al. (2004) discussed several possible implementations of the system, varying both classifiers and feature sets. Their best result, averaged over the scores reported for their six classes, is around 86.66% (precision said to be above 80% for some classes).[16] Finally, Leuski (2005) reported an average precision of 87% and average recall of 82% over four classes, but his approach treated entire messages as a single instance rather than individual utterances.

Our system's performance—overall accuracy 79%, average precision and recall 84.6% and 81.7%, respectively—falls within the same range as these results. These numbers suggest that there is still margin for improvement in this task, although it is unlikely that levels of performance similar to, for example, POS-tagging (97% accuracy) may be achieved: Pragmatic interpretation is often less clear-cut than assigning a part of speech to a single word, as seen also by the levels of human agreement on this task in our research and in other related work (e.g., Lampert et al. 2008 reported kappa of 0.54–0.79; Cohen et al. 2004 reported kappa of 0.72–0.83). However, we must also view our own results in the wider context of the application that it

is ultimately intended to be a component of, namely answer scoring. Further testing is needed to determine whether an accuracy rate of 79% is sufficient for this task, or whether better performance is needed to ensure the correct judgments about scores are made.

Discussion and Error Analysis

The 158 misclassified instances of this set were manually inspected to understand the possible causes of the classifier errors and whether they could be due to serious flaws in the system design, to issues related to the annotation scheme, or more simply to other factors such as misspellings and language errors in the input. Table 8 presents a confusion matrix, highlighting the source of the most frequent misclassifications. Rows refer to the correct classes, and columns refer to the classifier's decisions. For example, reading along the DD column, we find that 118 DD instances have been correctly classified and that 7 AA instances have been incorrectly labeled as DD.

Table 8

Confusion Matrix for L2 Data

Correct classes	Classifier's decision						
	AA	DD	DE	OT	QA	QQ	Total
AA	123	7	4	9	0	0	143
DD	2	118	38	3	0	1	162
DE	7	30	253	6	0	1	297
OT	2	2	8	94	0	1	107
QA	2	1	0	0	45	16	64
QQ	0	7	4	5	2	131	149
Total	136	165	307	117	47	150	

Note. AA = advisement, DD = disclosure, DE = commitment/factual statement, OT = other, QA = advisement in question form, QQ = simple question.

Overall, the findings of this analysis were very positive, as it revealed that many of the errors made by the classifier seem to be caused by difficulties in recognizing the subtlety of indirect speech acts. The misclassified instances proved problematic for the human annotators, too: 48% of these instances had not been assigned to the same class by all annotators. It may prove useful to discuss these instances with the annotators further to better understand the range of linguistic cues used by humans in interpreting speech acts. This information can in turn be

used to improve our model. In this section, we discuss the main error trends encountered and suggest some possible solutions. We will focus on requests, questions, and first person statements since misclassified third person statements only make up less than 10% of all errors.

Indirect Speech Acts

Almost all the misclassified requests, both AA (65%) and QA (84%), turn out to be noncanonical, indirect speech acts. They make up around 18% of all errors. As these can be hard to recognize for human speakers too—both native[17] and non-native—this finding is not surprising. In fact, some of these misclassified utterances are so indirect that it is debatable whether they have been correctly tagged in the first place. As an example of an indirect request/order, consider the following (all examples come from the dataset under discussion).

> I would like to meet you on Wednesday to discuss about the new building security system.
>
> [implied request: can we meet on Wednesday?]
>
> Original tag – AA (advisement)
>
> Classifier tag – DE (commitment/verifiable statement)

Although this is a very polite and indirect way of requesting a meeting, it does not contain any of the hallmarks of a typical request; in fact, the focus is on the speaker rather than the hearer, typical of politeness principles that seek to minimize impositions on the latter.[18]

The pattern of couching orders or request in indirect terms by focusing on the speaker rather than the hearer can be also seen in this example:

> When you go to the office on Monday, I recommend that you use the subway.
>
> [implied order: take the subway]
>
> Original tag – AA (advisement)
>
> Classifier tag – DD (disclosure of feeling)

In fact, the reactions of some native speakers show there is disagreement about the nature of this utterance; some do not consider it to be an order at all, but merely an opinion of the speaker. In this case, the classifier's choice of DD is indeed more appropriate than the annotation of AA given to it.

The misclassified QA instances present similar characteristics; many of them are a variation on the following example:

Is there anything that I have to prepare for the work?

[implied request: tell me what I have to do]

Original tag—QA (advisement as question)

Classifier tag—QQ (simple question)

The classifier is not wrong in recognizing these as questions. However, the human annotators agreed that this kind of question meets one of the key criteria for inclusion in the QA class, namely that it requires some action on the part of the hearer (in this case, giving the speaker a list of tasks, for instance), though there are no obvious cues signaling this.

In the sentences mentioned above, which are typical of this set of errors, a disconnection is present between the surface form of the utterance and its intended meaning. Clearly the classifier is basing its predictions on a set of surface features, as its incorrect class assignments do fit the surface features of these utterances. Recognizing indirect speech acts remains a major obstacle for this task, as some of the factors that force a particular interpretation of an utterance are extralinguistic, for example social distance between participants or knowledge of the physical context surrounding them. Indeed, Lampert et al. (2006) acknowledge these difficulties by excluding indirect speech acts from their analysis altogether and focusing on direct ones only. If we did the same with our corpus, our accuracy would rise to 86%. This solution cannot be considered workable, however, as indirect speech acts do occur in everyday life and must be accounted for. One possible approach regards the use of probability distributions, which will be discussed in greater detail below. Further study of the nature of indirect acts—at least the most typical examples—is also needed to establish whether shared characteristics among could be represented as features, too. Perhaps this problem could be overcome if a wider range of unigrams were used.

First Person Statements

Misclassified DDs (disclosures of feeling) and DEs (verifiable first person statements) make up 55% of all the classifier errors. The analysis of these instances brings to the forefront two possible issues with the system's development regarding the taxonomy chosen and the use of

modal verbs as features. Regarding the latter, we observe first of all that over one-third of DD instances classified as DE happen to include a feature typical of DE, either a modal verb or a am+present participle structure (or both), for example:

I can't help wondering what is going on.

I'm sure I will succeed in doing that work.

I'm wondering I will be able to work with him. I'm doing great!

These are not commitments or statements of facts, as they describe the speaker's state of mind. While it is important that the classifier has learned to associate modals to commitments—since they do most often occur there—they do also appear in other speech acts, and this overreliance on one feature must be addressed if performance is to improve. This issue also occurs in the other direction; that is, if we observe the misclassified instances of DE, we find that almost all of them do not have a modal verb. For example:

I have an appointment in the evening with my friend.

I have two years of experience working there.

I have some work to do after that.

At this moment I have nothing to do.

It is immediately evident that these kinds of statements are somewhat different in structure from commitments. Although their misclassification as DD is wrong from the perspective of the current taxonomy—since one of the criteria for annotation as DE is that the utterance must contain verifiable information, which is the case here—it may not be entirely wrong conceptually. Indeed, uttering a fact and making a commitment are statements of a different nature and would not actually be grouped together in classical pragmatic speech act taxonomies (according to Searle, 1979, for example, the first would be a representative and the second a commissive), so assigning them to the same class may have been a misjudged decision on our part. It is interesting to see that the speech act model acquired is sufficiently sophisticated to pick up this discrepancy and favor the core type of DE in its class assignment, drawing lines of demarcation in language that may correspond more closely to its natural categories than the ones artificially imposed by us.

The effect of our annotation policy is also evident with regard to a set of misclassified QQs, over half of which suffer from a similar problem of being tagged as DD or sometimes DE instead of QQ. This classification stems from the fact that, in listing the criteria for inclusion in this class, we stipulated that indirect or embedded questions would also count as questions. However, the classifier does not always recognize them and tags them as DD or DE instead. Some examples of this are (classifier tag given in square brackets):

I need to know the exact location of the office. [DD]

I just wonder if I can get the paycheck for Monday. [DE]

I really want to know about what affair that I take. [DD]

Again, the classifier has correctly recognized that these utterances share many surface characteristics with typical DD or DE sentences, and it is presumably relying on these characteristics to make its class assignment. These examples are indeed rather indirect speech acts (and in fact often one of the annotators disagrees about their nature) and perhaps their inclusion in the QQ category should also be rethought.

From a more practical point of view, when revising the taxonomy it may be most useful to focus on the needs of the application in the context of answer scoring. For example, if it is important for the scoring model to know that a commitment is or isn't present, but the nature of the other kinds of first person statements is irrelevant, one could change the first person categories to commitment and everything else, abandoning the distinction between verifiable and nonverifiable statements and emotions. In fact, other researchers, notably Lampert and his colleagues (cf., for example, Lampert et al., 2008b), have found themselves significantly changing and simplifying their annotation scheme after an initial round of experimentation.

Of course, no scheme is likely to cover all the subtleties of the pragmatic aspect of language, and some cases will always elude a classifier. For example, around a fourth of the misclassified DD instances have been assigned classes that could, upon reflection, also be appropriate; the following are just a few examples (classifier tag in square brackets):

I'll do my best. [DE]

I hope you can come to the office before my appointment. [AA—also suggested by one of the annotators]

I am so happy that I can help you. [DE]

I'd like to help you. [DE]

One of the crucial issues underlying these misclassifications is that pragmatics relates to the *social* aspect of communication, which relies on features that go beyond the formal structure of an utterance (e.g., on knowledge of the relationships between the participants of a conversation, or of their shared background, which can help understand a given statement in the intended key, even if its form could suggest several other possible interpretations). This extralinguistic information is very difficult to include in a computational model of language. A step toward including contextual awareness could be the addition of features that record the tags of previous and subsequent utterances, but this would only account for a small component of context: for example, it still would not give us any information about the relationship between the participants of the conversation.

L2 (Second Language Acquisition) Errors

Contrary to expectations, only 17 errors (10%) can be wholly or partly ascribable to poor English, which is perhaps a reflection on the high scores obtained by these test-takers across the board. An example follows:

I can have some questions which I should talk with you.

Original tag – DD

Classifier tag – DE

Based on what we have observed previously on the strong link between modal verbs and the tag DE, it is likely that the spurious *can* in this sentence has misled the classifier into tagging it as a commitment of some kind. This effect may have been further reinforced by the presence of the second modal verb *should*, which is arguably also not used appropriately in this context.

Often the errors are as simple as wrong punctuation or capitalization, which is of particular relevance to an e-mail-based task as these mistakes are not necessarily specific to L2 writers but can occur easily when typing quickly, as one tends to do in e-mails. These errors can be sufficient to throw off the classifier, as in this example:

Could you give me the right time when I will be arrived there.

Original tag – QA (advisement as question)

Classifier tag – AA (advisement/request)

Presumably the lack of a question mark at the end is leading the classifier to tag it as AA rather than a question-request; however, this error is relatively minor, because the classifier is still recognizing the request nature of the utterance and is only mistaken as to its form.

Sometimes, a missing apostrophe or uppercase letter means the parser fails to recognize the subject correctly as *I*:

On monday, Im scheduled to vacate.

Original tag – DE

Classifier tag – OT (other)

Here, the missing apostrophe means the parser does not see the subject as the pronoun *I*, but as some unknown noun *Im*, so naturally the classifier cannot assign the instance to any first person class and defaults to a third person statement.

Although these kinds of errors only make up a small proportion of all misclassifications, it is important to find ways to account for them. Preprocessing of the text, in the form of spelling and grammar checking, is a possible solution and is planned for the coming year.

Multiple Outputs

As a general observation, in all the experiments run so far, we require the classifier to output one decision label only. We also have the alternative of obtaining a probability distribution for each instance, that is, the classifier can output the probability of the instance belonging to each of the possible classes. This information could be very useful for cases where there is, in fact, disagreement about class membership among human judges, too, and could assist in improving accuracy by, for example, allowing us to set a threshold where instances with two equally plausible scores would be skipped or flagged.

We have tested this approach initially just on the set of misclassified instances. There are two ways in which we can rely on the information given by the probability distribution output: the *absolute approach* and the *relative approach*. In the absolute approach, we discard any instances where the probability of the top-ranked class is lower than a certain threshold—for example, 0.75. This means that the classifier would only output a classification for cases where

the instances have a 75% probability or higher of belonging to the given class. When applying this threshold, we find that almost all the misclassified instances are filtered out: the error rate decreases by 82%, and the new overall accuracy is 96.59%. This figure is very high, but this approach does mean that over 14% are not classified at all. Among the skipped instances are a large number of the indirect requests. In the example below, which was already introduced above, we can see that the probability of it belonging to any one class is not above 25% (original tag: AA).

When you go to the office on Monday, I recommend that you use the subway.

[implied order: take the subway]

AA 0.100 **DD 0.254** DE 0.166 OT 0.234 QA 0.014 QQ 0.230

The relative approach, on the other hand, focuses not on the actual score of each class but on comparing the classes' probabilities to each other. For example, we stipulate that the classifier must skip instances where the difference between the top two ranked classes is within a certain margin, such as 0.1. We find upon inspecting the data that one-quarter of misclassified instances fall into this category; that is, the correct class assignment is the second ranked choice, and it is within 0.1 of the first ranked choice. Applying this filter would raise accuracy to 87%. A more stringent one, requiring the gap to be 0.05 or less, correspondingly brings a smaller improvement in accuracy to 84.92%. This kind of threshold is expected to filter out instances that could belong to more than one class and might require wider knowledge to be assigned to the correct one in a given context. Most of the instances covered by this filter belong to the DE (verifiable first person statement) class, in particular, utterances that have been classified as DD (disclosure of feeling) instead (e.g., *At the moment I have nothing to do*). The small margin of difference between the DD and DE classes mirrors the intuition that the taxonomy we have chosen may not be dividing data into classes that tend to naturally cluster together.

The relative approach brings a smaller increase in accuracy than the absolute approach, but it does have the advantage of attempting a classification for every instance, thus not affecting coverage yet still reducing the amount of errors by 25%. Furthermore, in light of the observations made above with regard to the difficulty of assigning labels to certain speech acts and in the context of a scoring task, it may be of greater usefulness to output the top two choices

rather than skip cases altogether. The top two choices will likely correspond to two possible interpretations of the utterance and could then be flagged for inspection by a human scorer.

Further Testing: Novel Data

Performance on Other L2 (Second Language Acquisition) Data

The real measure of the system's success—and of its potential usefulness for ETS—lies in assessing its performance on data taken from a different set than that on which it was trained. This change is the only way to measure whether the models of speech acts it has acquired are general enough to be used for different prompts or applications. To establish this, we tested the model on three different kinds of datasets: L2 data from a different prompt, comparable L1 data (which shall be referred to as TOEIC L1), and e-mail data of a different nature from the Enron dataset.

The other L2 data comes from a different TOEIC administration, and therefore a different pool of test-takers. This measure ensures that no external factors, such as having text written by the same learners, bias this test of the system's performance. The conditions are comparable to those that we would have if the system were normally used for answer scoring, with no previous knowledge of the pool of test takers under examination.

The dataset consists of 948 utterances taken from e-mails written in response to a different prompt, namely an inquiry regarding a disputed bill. Overall accuracy is not as high for this set: 67.93%. For comparison, a majority baseline (always selecting DE, first person verifiable statement) would give 24.5% accuracy; using the ngram+first and last word features only gives 54.11% accuracy, while adding punctuation, too, brings average accuracy to 56.4%; the BOW baseline is 57.70%. The BOW baseline is much lower than the full-featured set accuracy figure, a trend that we will also observe with the L1 data in a later section. These findings suggest that the level of abstraction introduced by syntax-based features enables better generalization to novel data. Although good performance with a simple BOW model is possible, these figures show that an approach centered entirely around the lexicon, with no consideration for the more abstract structural properties of speech acts, is likely to be heavily dependent on the topic(s) discussed in a particular dataset and therefore less easily applicable to a novel dataset.

Returning to the full-featured set, both precision and recall are negatively affected, giving average scores of 74.20% and 70.68% respectively, a difference of 10%. Although all classes are affected in some measure, the figures for DE stand out in particular: precision 63.24%, recall

48.28%. Indeed, misclassified instances of DE (those that are DE but have not been recognized as such by the classifier) make up nearly 40% of all errors of this dataset, a greater proportion than previously observed.

It is therefore crucial to understand what may be causing this issue. Almost 70% of the misclassified DE instances have been incorrectly classified as OT (other), which is also somewhat unusual: typical confusion patterns are between the two first person classes, DD (expression of feeling) and DE, instead. Closer inspection of these instances reveals that almost all of them share one trait: the subject is in the first plural rather than the first singular person (e.g., *We'll check if there is some error with the payment*). In the training data, there were few or no instances of the first plural due to the nature of the topic, which involved an agreement between two individuals. In this test item, on the other hand, the writer is responding on behalf of his organization or department, which makes the use of the plural a plausible choice. However, the lack of commitments in the first person plural in the training data means that the classifier we have developed is not able to assign the correct class to these utterances. If we were to disregard this set of errors, to ensure a fairer comparison between datasets, accuracy for the L2 novel data rises to 76.48%, only 6% less than the accuracy observed for the data taken from the same pool as the training data. Furthermore, this problem also surfaces in the DD instances, in cases such as *We are sorry for this delay*. These instances constitute a smaller proportion of all errors (22.4% of the total); nevertheless, it is clear that overcoming the *we* vs. *I* issue, as will be discussed below, is likely to remove a major obstacle for the classifier.

Performance on L1 (First Language Acquisition) Data

From the discussion above, it seems that the main obstacle to the system's wider applicability is not a change in the general topic but rather the presence of different ways to formulate the same speech act. By testing the system on L1 data, too, we can confirm whether this is indeed the case or if further issues need to be addressed. In particular, we anticipated above the noticeable structural and lexical differences between native and non-native e-mails. We can now determine whether these differences negatively affect the system's performance.

The TOEIC L1 data also consists of e-mails. It has been collected from TOEIC Reading test exam scripts, as some of the TOEIC Reading tasks require the test-taker to read the text of an e-mail and answer questions designed to test his or her comprehension of it. These texts form an interesting comparison corpus as they are written by native speakers. Although they are not real,

being written for the explicit purposes of the tests, they aim to be as realistic as possible in portraying a wide variety of situations encountered in everyday work life.[19] The dataset consists of 1,721 utterances (200 e-mails). As well as being on a different topic from the training set e-mails, they do not form a homogeneous group in that all the e-mails differ from each other in content because they come from several test scripts rather than a single group of test answers. Topics of these e-mails include requests for meeting and appointments, clarifications about business transactions, requests for information, and orders.

Accuracy on this set is lower than that of the L2 data: 65.19% (precision 62.41%, 59.30%; baseline with just ngrams + first +last word 54.27% accuracy, baseline of ngrams + first _ last word + punctuation 57.66% accuracy; BOW baseline 57.4%). This marked difference from the L2 results suggests that the shift from non-native to native language may play an important role; some possible indications of how to overcome these issues are addressed later in this report. The most striking figure comes again from a first person class, this time DD, which obtains recall of just over 25%. QA and QQ also have low recall, at around 55%, but both classes are very small in this dataset (52 and 20 respectively), which might distort their scores. The misclassified DD instances are being assigned not just to the DE class, as one might expect, but also in large numbers to the AA and OT classes.

While the latter is easily explained by noting that the classifier assigns the OT label to instances that have *we* as a subject, as discussed above, the former is more surprising. Manual inspection of these misclassified instances reveals that the reason for the erroneous AA label is in all likelihood overfitting of a pattern observed in the training data, namely the use of the phrase *I am looking/I look forward to*. In the L2 data, this situation occurs almost exclusively in utterances such as *I look forward to your prompt reply*, which are coded as AA, as they are interpreted as polite and indirect requests for a speedy answer on the recipient's part. In the TOEIC L1 data, however, the phrase also occurs very often in utterances that are in fact conveying the speaker's thoughts or opinions:

I am looking forward to showing your employees our software.

I look forward to working with you.

Although this construction does not appear in the L2 data, it is not a very unusual one, and it is possible that non-native speakers may choose to use it, too. Testing our system against new

sources of data proves useful in highlighting areas where overfitting to the training data might be an issue.

Regarding the confusion between DD and DE, many of the points raised in the discussion of the L2 data also hold for this dataset. In particular, we noted above that the presence of modal verbs in DD statements can mislead the classifier into assigning the utterance a DE tag instead. This issue comes to the forefront with a subset of instances using the phrase *I would like*:

I would like to express my thanks.

I would like to discuss some issues with you.

I would like to see the report you completed.

These kinds of statements are rare in the training data, which might explain why the classifier relies on the presence of the modal verb *would* to select the DE class. They have been annotated as DD because in a basic way they represent the thoughts of the speaker: We do not have a way of verifying whether they really do want to discuss issues or see the report. However, it can be argued that the true function of these statements is not just to convey the speaker's feelings, but to actually perform a further act: in the first case, thanking, and in the other two cases, conveying a request in a very polite manner. We have already observed earlier that one of the central principles of politeness seeks to minimize impositions on the hearer: By phrasing these requests in terms of his or her own perspective rather than explicitly mentioning the hearer, the speaker is achieving this effect. In particular, requests formulated in this indirect way make it easier for the hearer to respond, or even evade the request if desired, especially in case of a negative answer: Because they are not explicit questions or orders, no explicit response is required. As an example, compare the following two exchanges; in the second one, both speaker and hearer avoid face-threatening situations:

S: Can I see the report you completed?

H: No, sorry, it's confidential.

S: I would like to see the report you completed.

H: Oh, actually that report is confidential.

The implications of these differences between L1 and L2 data will be discussed further below.

Finally, we also make use of the Enron corpus, introduced above, since it represents a type of text that is not only written by native speakers but is also real rather than just realistic. Although the main focus of our application is on the test-taking domain, meaning that it is unlikely to be used on real-world data, assessing it against this kind of data gives a measure of its wider applicability and can shed further light on any differences between real-world e-mails and the test answers.

Our Enron data consists of 1,396 utterances taken from the 250 e-mails annotated for bigram extraction. Accuracy on this data is 70.27% (precision 69.58%, recall 65.5%; baseline for ngrams + first+ last word 49.3%, ngrams + first + last word + punctuation 51.6%; BOW baseline 60.2%). Yet again, the main classes affected are the first person ones, DD and DE, in particular the former, with recall of just 34% and precision of 55%. These results further confirm that although at the current state of development our system has the potential to correctly identify speech acts across a variety of corpora, further work, in particular on the first person classes, is needed to ensure wider applicability.

Furthermore, we note that accuracy on the Enron data is higher than on the L1 TOEIC data. This result may appear surprising in some ways, since one might have expected the L2 data used in developing the system to be more similar to the TOEIC L1 data than Enron: even allowing for differences in content, the wider context of a learner-friendly text type would support that belief. We hypothesized that this result might be due to the fact that using the same Enron data for this evaluation as we did for bigram extraction biases the evaluation, since in a way the Enron data is not entirely novel for the classifier. To test this assumption, we also evaluated the datasets without the bigram feature; but the difference remains: L1 TOEIC accuracy is almost unchanged 65.48% and Enron accuracy is 68.69%. Perhaps this finding is due to the fact that the Enron e-mails are more closely clustered around the topic of office life, as are the L2 data, while the TOEIC data, while also pertaining to workplace issues, cover a wider spectrum of topics.

The results of testing the classifier on novel sources of data show that the principles at the core of our system are valid, as a large proportion of data from a variety of corpora is correctly classified. They have also highlighted areas where further work is needed to ensure wider applicability and avoid overfitting to the data used in the first phase of development and training.

The analyses conducted so far suggest that the issues identified can be addressed with relative ease. For example, the problem related to the presence of first person singular and plural pronouns can be solved by modifying the subject feature so it records only the person of the subject and not the number, then all commitments would share the trait of having a first person subject, regardless of whether the pronoun is singular or plural.

Other issues arising from the wider range of syntactic variety encountered in the other corpora—and the fact that similar phrases are used in different speech acts (cf. for example *look forward to*)—can be dealt with by including a wider range of data in training the system. However, we need to maintain a balance between developing a model that can generalize easily and keeping in focus the main goal of the application, which is to work with L2 rather than L1 data. It would be counterproductive to train the classifier to recognize instances of language that it is unlikely to encounter in the context of test scoring. The ideal approach would rely on a broader range of L2 data, especially those taken from higher-scoring test responses, to ensure we have a model of language that conforms to what scorers expect of the learners.

Differences Between Native and Non-Native E-Mails

The discussion above would appear to suggest that what constitutes a correct answer for the TOEIC task is not necessarily the same thing as a typical, well-formed L1 e-mail. Indeed, many differences among the corpora have already been highlighted in the discussion of the classifier's performance. What follows are only preliminary considerations on the topic. More in-depth analysis is needed, such as a feature-by-feature comparison of the corpora. We can begin by looking at the distribution of the various speech acts across the corpus. Although the comparison is only indirect due to the different nature of the content—we do not have a collection of L1 e-mails written in response to this particular test prompt—some general observations are still possible. Table 9 reports the frequency of each class within the corpus (L2, L1, and Enron e-mails) as a percentage of the total. The L2 data shares some characteristics with each of the two L1 corpora. Requests, for example, are present with roughly the same frequency in all three datasets. The DD and DE classes are also similar in the two TOEIC corpora, while they are much less frequent in the Enron data. It is difficult to speculate as to the reasons for this finding, which may lie in extralinguistic factors: for example, we do not know what the e-mail culture of the Enron workplace was like and whether it was considered more appropriate to express opinions and commitments in person or over the phone rather than over e-mail.[20] It is

also possible that some of these differences arise from the fact that the TOEIC e-mails (both L1 and L2) have been created in response to testing purposes and, while they aim to be realistic, they may not necessarily be characteristic of e-mail communication in the workplace.

Table 9

Distribution of the Six Classes in L1 (First Language) and L2 (Second Language) Data

Class	L2	L1	Enron
AA	15.51%	16.39%	18.48%
DD	17.57%	14.99%	9.81%
DE	32.21%	29.87%	21.42%
OT	11.61%	34.57%	35.67%
QA	6.94%	3.02%	6.16%
QQ	16.16%	1.16%	8.45%

Note. AA = advisement, DD = disclosure, DE = commitment/factual statement, OT = other, QA = advisement in question form, QQ = simple question.

More striking differences are found in looking at the patterns for the OT class and both kinds of questions. The large number of QQs in the L2 data may be an artifact of the corpus used, as the rubric for that test item explicitly requires students to ask two questions, but it is similar in the other L2 data we have annotated, too. It would seem that native speakers prefer not to ask too many direct questions in their e-mails; this claim requires further investigation, but it would be in agreement with our previous observations on the nature of requests and a tendency to prefer indirect to direct requests, in line with the face-saving principles outlined above.[21]

The greater share of OT statements was already anticipated when we noted that L1 e-mails tend to include a larger amount of supporting information and background details; that is, the kinds of statements that belong to the OT class.[22] The absence of such statements in many of the learner e-mails might inadvertently provoke a negative reaction in a native speaker recipient. They might be perceived as being somehow incomplete despite being formally correct: perhaps too direct and brusque, too quick to get to the point, even rude. Compare the following two (fictitious) examples:

Dear John,
Can you come to a meeting with me tomorrow? It’s at 10.
Mary

Dear John,
There is a meeting at 10 tomorrow with our manager, and I think it would be good if you attended. You have a better grasp of the figures than I do, and he will be impressed to hear of your work directly from you. Could you come along?
Thanks,
Mary

In the second example, Mary gives some context for her request, so that it not only comes across as less brusque, but it also receives some motivation: having background information can help John make an informed decision about the meeting. Of course, once again extralinguistic factors also come into play. The power and social relations between speaker and hearer greatly affect these considerations. For example, in American and many European contexts, it is generally accepted that a request coming from a more senior person need not be accompanied by much explanation. Conversely, it would seem very inappropriate for a junior worker to issue a request like the first example to his or her superior.

A further line of inquiry might also consider the EFL speakers' stylistic choices in the context of research on stance, which is the way "speakers and writers can express personal feelings, attitudes, value judgments or assessments" (Richards & Schmidt, 2002, p. 507). Stance is conveyed, among other means, through many of the elements typical of speech acts: modal verbs, pronoun choice, hedges (adverbs or phrases qualifying one's statement, such as *maybe* or *I think*), and verbs expressing desires (*want, wish, hope, would like*). A strong link exists between stance and social/professional identity (for an up-to-date overview, cf. the papers collected in Jaffe, 2009), so it is important that non-native speakers be aware of how their stylistic choices can affect the way they are perceived in the workplace.

We therefore return to the question posed at the start of this discussion: How do the test e-mails differ from native e-mails, and does it matter? Many considerations bear on this issue. Time constraints are a central one: during the test, students have a limited amount of time in which to compose their answer. It is natural that they will prioritize the use of the speech acts required by the rubric and only include other material—such as the supporting statements described above—if time allows. Perhaps learners are not in fact averse to using more OT

statements: But this claim can only be tested if we had access to real-world e-mail data, which is very hard to obtain.

The TOEIC Writing test aims to prepare students to work in the global workplace: The focus is not necessarily on being as native-like as possible, but on achieving successful communication, which means both making oneself understood and avoiding giving offense. In this respect, then, it does not matter if the e-mails do not conform to a native model, as long as their message is clear. However, we have also seen that the simple presence or absence of a linguistic structure in an e-mail can in fact deeply affect the perceived politeness in ways that are important to be aware of as a learner. It is unlikely that one can be faulted for being too polite, but the opposite can have undesirable consequences, especially when workplace matters are at stake. Our view therefore is that awareness of the linguistic realizations of politeness principles must be included in L2 instruction, although the present analysis has only grazed the surface of a very rich and constantly evolving field.

Conclusion

This report discussed several issues of relevance for the TOEIC Writing e-mail task. We described an approach to the automated scoring of this task, focusing on the presence of speech acts in the test responses. The computational model for automated speech act identification we developed achieves up to 79.28% accuracy; we have suggested possible solutions to achieve better performance. We also compared our TOEIC e-mail data to corpora of speech-act annotated native English e-mails, and discussed the impact of differences in speech act use between native and non-native English. We believe this study is a useful first attempt at developing a comprehensive approach to the automated scoring of the TOEIC e-mail task.

References

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, *4*(3). Retrieved from http://escholarship.bc.edu/ojs/index.php/jtla/article/viewFile/1650/1492

Attali, Y., Powers, D., Freedman, M., Harrison, M., & Obetz, S. (2008). *Automated scoring of short-answer open-ended GRE subject test items.* Princeton, NJ: ETS.

Austin, J. L. (1962). *How to do things with words.* Oxford, UK: Clarendon Press.

Bardovi-Harlig, K., & Dornyei, Z. (1998). Do language learners recognize pragmatic violations? Pragmatic versus grammatical awareness in instructed L2 learning. *TESOL Quarterly*, *32*(2), 233–262.

Bejar, I. (in press). *Can speech technology improve assessment and learning? New capabilities may facilitate assessment innovations*. Princeton, NJ: ETS.

Berry, M., Browne, M., & Signer, B. (2007). *2001 topic annotated Enron e-mail data set.* Philadelphia, PA: Linguistic Data Consortium.

Biber, D. (1988). *Variation across speech and writing.* Cambridge, UK: Cambridge University Press.

Biesenbach-Lucas, S. (2005). Communication topics and strategies in e-mail consultation: Comparison between American and international university students. *Language Learning & Technology*, *9*(2), 24-46.

Bird, S., & Loper, E. (2004). *NLTK: The natural language toolkit. In The companion volume to the proceedings of the 42nd annual meeting of the Association for Computational Linguistics*. Retrieved from http://aclweb.org/anthology-new/P/P04/P04-3031.pdf

Brown, P., & Levinson, S. (1987). *Politeness: some universals in language usage.* Cambridge, UK: Cambridge University Press.

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion online essay evaluation: an application for automated evaluation of student essays. In *Proceedings of the fifteenth annual conference on innovative applications of artificial intelligence*. Retrieved from http://www.ets.org/Media/Research/pdf/erater_iaai03_burstein.pdf

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

Carvalho, V., & Cohen, W. (2005). *On the collective classification of e-mail "speech acts." In Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.4017&rep=rep1&type=pdf

Carvalho, V., & Cohen, W. (2006). Improving e-mail speech act analysis via n-gram selection. In *Proceedings of the HLT-NAACL analyzing conversations in text and speech workshop*. Retrieved from http://www.aclweb.org/anthology/W/W06/W06-3406.pdf

Clark, S., & Curran, J. (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics, 33*(4), 493–552.

Cohen, A. (2008). Teaching and assessing L2 pragmatics: What can we expect from learners? *Language Teaching, 41*(2), 213–235.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220.

Cohen, W., Carvalho, V., & Mitchell, T. (2004). *Learning to classify e-mail into "speech acts."* In *Proceedings of EMNLP 2004*. Retrieved from http://www.aclweb.org/anthology-new/W/W04/W04-3240.pdf

Corston-Oliver, S., Ringger, E., Gamon, M., & Campbell, R. (2004). *Task-focused summarization of e-mail.* In M. Moens & S. Szpakowicz (Eds.), *Text summarization branches out: Proceedings of the ACL-04 workshop*. Retrieved from http://www.aclweb.org/anthology-new/W/W04/W04-1008.pdf

Cruse, A. (2000). *Meaning in language: An introduction to semantics and pragmatics*. Oxford, UK: Oxford University Press.

Curran, J., & Clark, S. (2003). *Language independent NER using a maximum entropy tagger. In Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003* (pp. 164–167), Edmonton, Canada. Retrieved from http://aclweb.org/anthology/W/W03/W03-0424.pdf

Curran, J., Clark, S., & Bos, J. (2007). *Linguistically motivated large-scale NLP with C&C and Boxer. In Proceedings of the 45th annual meeting of the Association for Computational Linguistics Companion volume proceedings of the demo and poster sessions* (pp. 33–36), Prague, Czech Republic. Retrieved from http://aclweb.org/anthology-new/P/P07/P07-2009.pdf

De Felice, R. (2009). *I'm afraid that I should say I can't: Negotiating refusals in L2 Business English e-mails.* Paper presented at the 5th Conference on intercultural rhetoric and discourse, Ann Arbor, MI.

De Felice, R., & Pulman, S. (2008). *A classifier-based approach to preposition and determiner error correction in L2 English. In Proceedings of the 22nd international conference on computational linguistics (Coling 2008).* Retrieved from http://www.aclweb.org/anthology/C08-1022

Deane, P. (2006). Strategies for evidence identification through linguistic assessment of textual responses. In D. Williamson, R. Mislevy, & I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 313–372). Mahwah, NJ: Lawrence Erlbaum.

Dornyei, Z. (1995). On the teachability of communication strategies. *TESOL Quarterly*, *29*(1), 55–85.

Dredze, M., Brooks, T., Carroll, J., Magarick, J., Blitzer, J., & Pereira, F. (2008). *Intelligent e-mail: reply and attachment prediction.* In *Proceedings of the 13th international conference on intelligent user interfaces*. Retrieved from http://doi.acm.org/10.1145/1378773.1378820

Dredze, M., Wallach, H., Puller, D., Brooks, T., Carroll, J., Magarick, J., ….Pereira, F. (2008). *Intelligent e-mail: Aiding users with AI.*In D. Fox & C. P. Gomes (Chairs), *Proceedings of the twenty-third AAI conference on artificial intelligence* (pp. 1524–1527). Retrieved from http://www.aaai.org/Library/AAAI/aaai08contents.php

ETS. (2010). The TOEIC user guide—Speaking & writing. Princeton, NJ: Author.

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W., Belenko, D., Vanderwende, L. (2008). *Using contextual speller techniques and language modeling for ESL error correction.* In *Proceedings of IJCNLP* (Vol. 51, pp. 449-456). Retrieved from http://research.microsoft.com/pubs/69477/IJCNLP.pdf

Georgila, K., Lemon, O., Henderson, J., & Moore, J. (2009). Automatic annotation of context and speech acts for dialogue corpora. *Natural Language Engineering*, *15*(3), 315–353.

Goldstein, J., & Sabin, R. (2006,). Using speech acts to categorize e-mail and identify e-mail genres. In *Proceedings of the 39th annual Hawaii international conference on system sciences – Volume 3*, Retrieved from http://dl.acm.org/citation.cfm?id=1109711.1109727

Jaffe, A. (Ed.). (2009). *Stance: Sociolinguistic perspectives.* Oxford, UK: Oxford University Press.

Jaszczolt, K. M. (2002). Semantics and pragmatics: meaning in language and discourse. London, UK: Longman.

Kaplan, R. (1966). Cultural thought patterns in inter-cultural education. *Language Learning*, *16*, 1–20.

Kasper, G. (1984). Pragmatic comprehension in learner-native speaker discourse. *Language Learning*, *34*(4), 1–20.

Kasper, G. (2001). Four perspectives on L2 pragmatic development. *Applied Linguistics*, *22*(4), 502–530.

Kasper, G., & Rose, K. (1999). Pragmatics and SLA. *Annual Review of Applied Linguistics*, *19*, 81–104.

Khosravi, H., & Wilks, Y. (1999). Routing e-mail automatically by purpose not topic. *Natural Language Engineering*, *5*(3), 237–250.

Koike, D. A. (1989). Pragmatic competence and adult L2 acquisition: Speech acts in interlanguage. *The Modern Language Journal*, *73*(3), 279–289.

Lampert, A., Dale, R., & Paris, C. (2006). *Classifying speech acts using verbal response modes.* In *Proceedings of the 2006 Australasian language technology workshop (ALTW)*. Retrieved from http://www.ict.csiro.au/staff/andrew.lampert/writing/papers/speechactsvrm-altw2006-lampert.pdf

Lampert, A., Dale, R., & Paris, C. (2008a. *The nature of requests and commitments in e-mail messages.* In *Proceedings of the AAAI workshop on enhanced messaging*. Retrieved from https://www.aaai.org/Papers/Workshops/2008/WS-08-04/WS08-04-008.pdf

Lampert, A., Dale, R., & Paris, C. (2008b). *Requests and commitments in e-mail are more complex than you think: Eight reasons to be cautious. Proceedings of the 2008*

Australasian language technology workshop (ALTW)., Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.2254&rep=rep1&type=pdf

Lampert, A., Paris, C., & Dale, R. (2007). *Can requests-for-action and commitments-to-act be reliably identified in e-mail messages? In Proceedings of the 12th Australasian Document Computing Symposium*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.153.9470&rep=rep1&type=pdf

Leacock, C., & Chodorow, M. (2003). c-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389–405.

Lee, J. (2009). *Automatic correction of grammatical errors in non-native English text.* Unpublished dissertation, MIT, Cambridge, MA.

Leuski, A. (2005). *Context features in e-mail archives.* In *Proceedings of the ACM SIGIR 2005workshop on information retrieval in context* (IRiX. Retrieved from citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.8687&rep=rep1&type=pdf#page=54

Levinson, S. (1983). *Pragmatics.* Cambridge, UK: Cambridge University Press.

Manning, C., & Schuetze, H. (1999). *Foundations of statistical language processing.* Cambridge, MA: MIT Press.

Meunier, F. (1998). Computer tools for the analysis of learner corpora. In S. Granger (Ed.), *Learner English on computer* (pp. 19–37). London, UK: Longman.

Mey, J. (2004). Between culture and pragmatics: Scylla and Charybdis? The precarious condition of intercultural pragmatics. *Intercultural Pragmatics*, *1*(1), 27–48.

Mildinhall, J., & Noyes, J. (2008). *Toward a stochastic speech act model of e-mail behavior.* In *Proceedings of the conference on e-mail and anti-spam (CEAS).* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.151.8042&rep=rep1&type=pdf

Minnen, G., Carroll, J., & Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering, 7*(3), 207–223.

Moon, K. (2002). *Speech act study: Differences between native and non-native speaker complaint strategies* (American University TESOL Working Papers 1). Retrieved from http://www.american.edu/cas/tesol/pdf/.../WP-2002-Moon-Speech-Act.pdf

Pulman, S., & Sukkarieh, J. (2005). *Automatic short answer marking. In Proceedings of the second workshop on building educational applications using NLP*. Retrieved from http://www.aclweb.org/anthology/W/W05/W05-0202

Ratnaparkhi, A. (1998). *Maximum entropy models for natural language ambiguity resolution.* Philadelphia: University of Pennsylvania.

Reppen, R., Ide, N., & Suderman, K. (2005). *American National Corpus (ANC) second release.* Philadelphia, PA: Linguistic Data Consortium.

Richards, J., & Schmidt, R. (2002). *Longman dictionary of language teaching and applied linguistics*. Harlow, UK: Longman.

Schmidt, R., & Richards, J. (1979). Speech acts and second language learning. *Applied Linguistics, 1*(2), 129–157.

Searle, J. R. (1969). *Speech acts.* Cambridge, UK: Cambridge University Press.

Searle, J. R. (1979). *Expression and meaning.* Cambridge, UK: Cambridge University Press.

Stiles, W. (1992). *Describing talk: A taxonomy of verbal response modes*. Thousand Oaks, CA: Sage.

Sukkarieh, J., & Bolge, E. (2008). Leveraging c-rater's automated scoring capability for providing instructional feedback for short constructed responses. In B. P. Woolf , E. Aimeur, R. Nkambou, & S. P. Lajoie (Eds.), *Lecture notes in computer science : Vol. 5091. Intelligent tutoring systems* (pp. 779–783). Berlin: Springer-Verlag.

Sukkarieh, J., & Pulman, S. (2005). *Information extraction and machine learning: Auto-marking short free text responses for science questions*. In *Proceedings of the 2005 conference on artificial intelligence in education: Supporting learning through intelligent and socially informed technology*. Retrieved from http://dl.acm.org/citation.cfm?id=1562524.1562609

Tetreault, J., Burstein, J., & De Felice, R. (Eds.). (2008). *Proceedings of the third workshop on innovative use of NLP for building educational applications.* Columbus, OH: Association for Computational Linguistics.

Tetreault, J., Burstein, J., & Leacock, C. (Eds.). (2009). *Proceedings of the fourth workshop on innovative use of NLP for building educational applications.* Boulder, CO: Association for Computational Linguistics.

Tetreault, J., & Chodorow, M. (2008). *The ups and downs of preposition error detection. In Proceedings of the 22nd international conference on computational linguistics* (Coling 2008). Retrieved from http://www.aclweb.org/anthology/C08-1109

Thomas, J. (1983). Cross-cultural pragmatic failure. *Applied Linguistics, 4*(2), 91–112.

Notes

[1] This is an issue for both written and spoken English, as set out for example by Bejar (in press).

[2] A related field is that of dialogue act classification: There is a large body of research on understanding intentions in dialogue modeling (a recent overview is in Georgila, Lemon, Henderson, & Moore, 2009), but the challenges are somewhat different as dialogue systems deal with synchronous rather than asynchronous communication, and with differently structured exchanges, for example, not as likely to contain complete sentences. Furthermore, these systems are usually designed with a limited information-seeking domain in mind, and as such the taxonomies designed for the annotation of dialogue acts are not appropriate for our task.

[3] We refer to utterances rather than sentences because longer sentences, with many coordinated clauses, are broken up at the annotation stage.

[4] In fact, the linguistics literature stresses repeatedly that the social context and the relationship between speaker and hearer are crucial components in the correct interpretation of a speech act (the same utterance said among friends may have a different value if said by a mother to a child, for example); however, the technical limitations of NLP analysis mean that we must restrict our focus to those surface linguistic features that are easily identifiable.

[5] http://svn.ask.it.usyd.edu.au/trac/candc/wiki

[6] To put this in context, we can compare the e-mails' type/token ratio to that of several kinds of written texts as reported by Biber (1988): official documents 47:8, academic prose 50:6, fiction 52:7, personal letters 52:5, professional letters 53:0; face to face conversations have a ratio of 46:1. However, this comparison is indirect as texts of less than 400 words are not analyzed.

[7] The calculations for the unigrams were obtained using NLTK (Bird & Loper, 2004), a suite of Python modules for NLP tasks available from www.nltk.org. No stemming is carried out, as we feel it is important to preserve differences in tense.

[8] For example, they could be selected on the basis of information gain instead.

[9] From the description in the LDC catalog: The 2001 Topic Annotated Enron E-Mail Data Set contains approximately 5,000 (4,936) e-mails from Enron Corporation (Enron) manually

indexed into 32 topics. It is a subset of the original Enron e-mail data set of 1.5 million e-mails that was posted on the Federal Energy Regulatory Commission website as a matter of public record during the investigation of Enron. The e-mail topics reflect the business activities and interests of Enron employees in that year. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T22

[10] N-gram counts from the corpus are freely available from http://www.americannationalcorpus.org/SecondRelease/frequency2.html

[11] It could be argued that this is not a wholly correct interpretation, as we are not carrying out a further comparison between workplace e-mail and personal e-mail. Therefore, we don't know if the characteristics that emerge are typical of e-mail in general or more specific to workplace e-mail. However, in the absence of available sources of nonworkplace e-mail, we rely on this simplified analysis as a first means of distinction.

[12] This terminology originates in the domain of information retrieval, where not all the data being tested is actually relevant—so it does make sense to speak of positive and negative, and true negatives are also needed—data that should not have been retrieved, and wasn't. In the context of this task, where all the data is relevant, and must be classified, the terminology may appear somewhat misleading.

[13] *Bag-of-words* refers to using all the words in the utterance as features without any further information about them.

[14] Although it could be argued that the first word feature and the bigrams indirectly represent some basic syntactical information, for example having a sentence start with a verb is a strong indicator of a question, and that the bigram set also includes sequences that have the subject-verb inversion of sentences (i.e., *can you* instead of *you can*).

[15] These findings are confirmed by feature subtraction experiments, whereby we remove a different feature each time for training and observe how its absence affects performance to better assess its contribution. While in many cases performance drops by 1% or less, the removal of the punctuation feature leads to a 2% drop in performance, and that of n-grams and of first and last word, 3.5% each.

[16] The precise figure is unavailable, as the results are actually reported in the form of error rates in a bar chart, and has been estimated by the present authors.

[17] The experience of not seeing one's polite and roundabout request acknowledged, or conversely appearing rude for not recognizing a similar request in someone else, is a common one.

[18] Brown and Levinson (1987), one of the core texts of the literature on politeness (which is too extensive to be treated here) introduced the concept of face-saving and face-threatening acts, stating that politeness principles seek to maximize the former and minimize the latter.

[19] We are very grateful to Trina Duke for suggesting this source of data and enabling access to it.

[20] We must remember also that the Enron e-mails date from 2000–2001, when perhaps e-mail use in the workplace followed different patterns.

[21] If a wider corpus investigation supports this claim regarding the use of requests in e-mails, one might ask whether these findings could have an effect on the TOEIC items to ensure the test's requirements are more reflective of real-life practices.

[22] Of course, this may also mean that in real or realistic e-mails there is a greater use of the medium to simply convey information, such as sending out announcements, confirmations of reservations, and so on.

Appendix A

Annotation Guidelines

For sentences containing conditionals, generally the emphasis should be on the content of the main clause and the tag should reflect that. So, in a sentence such as *If I send you the files now, can you reply by tomorrow?*, the tag should reflect the speech act represented by *can you reply by tomorrow*.

The annotation scheme considers two aspects of a speech act in its coding system, *form* and *intent* (roughly comparable to locutionary and illocutionary act). These could coincide, as in an order that sounds like an order—*Close the door!*—or not, as in an order which is formulated very indirectly—*I wonder if it would be quieter if we closed the door*. These two aspects are represented by the two letters in each tag, where by convention the first represents form and the second represents intent. Acts where form and intent coincide are called *pure*, and those where they don't are known as *mixed*. Three sample annotated e-mails are given at the end of these guidelines.

Acknowledgments (K)

These refer to the typical phrases one finds in the opening of e-mails, which usually acknowledge the previous e-mail or similar previous communication.

A pure acknowledgment (KK) is simply a one or two word phrase. This tag is used for a salutation, such as *Dear Tim* or *Hi Tim*, and also for a sentence at the beginning of the e-mails such as *Thank you for your e-mail*. Elements peculiar to e-mails, such as *To: Jim Jones*, can also be coded KK. Note that it is acceptable to have more than one KK element in a single e-mail.

However, one can also use stock phrases that have some content to them, in which case they will receive a mixed mode tag because they impart some thought while having the form of an acknowledgment. If the proposition expressed is something objectively verifiable, the code is KE, where E stands for edification (cf. below). For example: *I just received your e-mail.* If, on the other hand, the proposition expressed is related to personal thoughts or feelings, the code is KD, where D stands for disclosure (cf. below). An example is *I was very honored to receive your e-mail*, where the focus is on the speaker's thoughts, which we cannot verify for sure.

Closing (FW)

The FW (farewell) tag is applied to typical closing phrases such as *Sincerely, See you soon*, or the writer's name. Position within the text is a key factor here: These elements can only be tagged FW if they are actually at the end of the e-mail. The sentence *I hope to see you soon*, for example, is not an FW if it occurs at the beginning or in the middle of the e-mail.

A possible source of confusion here comes from phrases such as *Please let me know as soon as possible* or *I am looking forward to hearing from you soon*. While these utterances are often used as routine closing phrases, they do carry with them a presumption that the recipient will have to do something to respond to the writer's request. They are therefore coded AA for Advisements (cf. below) whether they are in the first or third person.

Questions (Q) and Advisements (A)

These two instances are treated together as they are often found in mixed mode. For something to be tagged as having a question form, it can either be a direct question, ending with a question mark, or be embedded in a declarative clause (e.g., *I want to know what time it is* or *I wonder how I much I can get for this job*).

Pure questions (QQ) are a genuine request for information that can be obtained without the hearer having to take special action (e.g., *What's your name? What time is the meeting?*)

Advisements can be formed as indirect questions or more explicit orders; their defining characteristic is that they request that the hearer do something. They attempt to affect the hearer's behavior, whether by requiring an action (*Please send me the files* or *I'd like to ask you to phone the client for me*) or a change in their mental state (*Don't worry about the meeting* or *You should feel more positive about the interview*). Second person pronouns feature heavily here. All of the above are examples of pure advisements (AA) since they are not in question form.

Often advisements can be phrased as direct questions for reasons of politeness. In this case they have the mixed mode (QA), as they have the form of questions but are actually intending to affect, or advise, the hearer's behavior (e.g., *Could you send me the files? Would you be able to drive us to the office? I was wondering if you could give me a call*). These utterances can be very tricky, as sometimes the author will use very roundabout ways to issue requests due to extreme politeness. In general, if the answer to the question requires any sort of action on the part of the recipient, it is to be considered an advisement. An example is *Will I be*

doing the same work as before?, which could be interpreted either as a straightforward question with a yes or no answer, or as a very indirect way of saying "Please tell me what I need to do." Generally, this latter interpretation is preferred for the purpose of this task.

Disclosure (D) and Edification (E)

These utterances are often hard to distinguish as they can occur in mixed mode, and the interaction between form and intent can be subtle. What follows are the basic principles for their identification.

Regarding form, disclosures are always in the first person while edifications are always in the third person/impersonal. This form reflects the crucial distinction that disclosures, as the name suggests, involve the sharing of the feelings or intentions of the speaker, while edifications involve the stating of objective information. A similar rule of thumb is that the truth of disclosures cannot be verified without having access to the speaker's thoughts, while the truth of edifications is objectively verifiable. Edifications provide data; disclosures are a revealing of one's self.

Some examples of pure disclosures (DD) are I am so happy to see you, I am so sorry to hear your news, or I have a question for you.

Similarly, some examples of pure edifications (EE) are The meeting is at 8 tomorrow or The figures for the year are ready.

However, things get complicated by the form vs. intent dichotomy, as something which has the form of a disclosure (i.e., be in the first person) may well involve objective information; and something which is in the edification form may impart some information which is subjective to the speaker.

Disclosures with an edification intent (DE) are easily verifiable (e.g., *I will bring the data* or *I will attend the meeting*. These have to be coded D for form because they are in the first person, but their intent is to convey actual information rather than express the speaker's feelings. The proposition they contain can be easily verified.

Conversely, edifications with a disclosure intent (ED) may refer to a third person, but their proposition may be impossible to verify without access to the speaker's private thoughts (e.g., *This really bothers me* or *Your decision does not make sense to me).* Although the code for form is E, as they are in the third person, what they convey cannot be objectively assessed.

Another tricky case in distinguishing between DE and DD concerns sentences where the speaker is making a commitment. For the purposes of this task, we believe that if one is committing to something, the fulfillment or otherwise of said commitment can be subsequently verified, and so these cases should be tagged DE. This criterion includes all sentences that have commitments hedged by the modal *can* (e.g., *I can finish my work before 8* or *I can take your place at the meeting*). Similar considerations apply to these sentences when they are embedded within another clause as the object of a verb such as *think*. Even though this situation represents a very cautious commitment, we decided it is a commitment nonetheless. For example: *I think I can attend the meeting on Monday*.

Other (OO)

If the coder feels that none of the existing tags are suitable, then OO may be used. This tag should occur very rarely (e.g., *You must be so upset right now*, or *Great! Good luck on your new job!*).

In general, if a sentence is a statement that refers to something that is not objectively verifiable but does not regard the speaker's thoughts or feelings, it should be coded as OO. Many of these will feature second person pronouns.

Uncodable (UU)

Finally, if the utterance is uncodable because of structural issues, for example, because the sentence is incomplete, UU must be used. As in the original guidelines, this tag can't be used simply when the tag is hard to determine, but only when the utterance is genuinely corrupt, as is occasionally the case in the data (e.g., the sentence is unfinished or impossible to understand). Examples include truncated utterances (*Can you gi*), fragments (*Before work on Monday.*) and malformed sentences (*So I have my job efficiently.*).

Example 1

Dear Laura,	KK
Thanks for your e-mail.	KK
I will be happy to take the seminar speaker for lunch on Friday.	DE
I'm sorry you have to miss it.	DD

Can you tell me where the speaker is staying?	QA
Will I need to pay for his lunch?	QQ
Thanks,	FW
Rachele	FW

Example 2

Dear Davis,	KK
How are you?	KK
You must be feeling stressed!	OO
Of course I can go to the meeting.	DE
It’s my pleasure.	ED
So what time should I go there?	QQ
And please tell me what to expect at the meeting.	AA
I hope you will e-mail me back asap.	AA
Yours,	FW
Andy	FW

Example 3

Hi Tom,	KK
I know about your problem.	KD
I think I am able to help with it.	DE
I’ve dealt with this before so it’s not hard for me.	DD
I need some more information though.	DD
So can you show me the figures?	QA
I look forward to hearing from you soon.	AA

Appendix B

List of Unigrams Used

Word	Class	Word	Class	Word	Class
Advise	AA	confident	DD	accept	DE
answers	AA	curious	DD	available	DE
appreciate(d)	AA	feeling	DD	checked	DE
Clearly	AA	forgot	DD	checking	DE
consider	AA	glad	DD	client(s)	DE
convenience	AA	happy	DD	heard	DE
earliest	AA	honor	DD	hearing	DE
Fast	AA	hope	DD	introduce	DE
Feel	AA	impression	DD	manage	DE
forward	AA	miss	DD	plans	DE
further	AA	perfect	DD	promise	DE
hesitate	AA	proud	DD	try	DE
letter	AA	sorry	DD	unfortunately	DE
looking	AA	totally	DD	willing	DE
message	AA	understand	DD		
notify	AA	understood	DD	asks	QA
please	AA	valuable	DD	borrow	QA
points	AA	wish	DD	direction	QA
prompt	AA	worried	DD	duties	QA
read	AA	worrying	DD	explain	QA
receiving	AA			instruction(s)	QA
reply	AA	interesting	OT	specifically	QA
respond	AA	looks	OT		
response	AA	luck	OT	does	QQ
wait(ing)	AA	seem(s)	OT	firstly	QQ
worry	AA	sounds	OT	secondly	QQ
write	AA			wonder(ing)	QQ
yourself	AA				